SIMULATION OF A WHOLE-CELL WITH THE MINIMUM NUMBER OF GENES NECESSARY FOR SUSTAINED REPLICATION

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Jordan Crawford Atlas

May 2010

SIMULATION OF A WHOLE-CELL WITH THE MINIMUM NUMBER OF GENES NECESSARY FOR SUSTAINED REPLICATION

Jordan Crawford Atlas, Ph.D.

Cornell University 2010

One important aim of synthetic biology is to develop a self-replicating biological system capable of performing useful tasks. A mathematical model of a synthetic organism would greatly enhance its value by providing a platform in which proposed modifications to the system could be rapidly prototyped and tested. Such a platform would allow the explicit connection of genomic sequence information to physiological predictions. As an initial step toward this aim, a Minimal Cell Model (MCM) has been formulated. The MCM is defined as a model of a hypothetical cell with the minimum number of genes necessary to grow and divide in an optimally supportive culture environment. It is chemically detailed in terms of genes and gene products, as well as physiologically complete in terms of bacterial cell processes like DNA replication and cell division.

A mathematical framework originally developed for modeling *Escherichia coli* has been used to build the platform MCM. To lay the foundation for designing an MCM, sensitivity analysis and event detection methods applicable to the *E. coli* model are presented. An updated version of the *E. coli* model that links detailed genomic information about the location of *dnaA* genes and DnaA binding sites on the chromosome to physiological predictions has been developed. The model suggests that the concentration of DnaA binding boxes on the chromosome is critical to determining cell growth and behavior. This

update is the first example of including detailed genomic information in a hybrid bacterial cell model, which was an important step toward the massive inclusion of new genes in the MCM.

An MCM with 241 product-coding genes (those which produce protein or stable RNA products) is presented. This set is genomically complete and codes for all the functions that a minimal chemoheterotrophic bacterium would require for sustained growth and division. It is shown for the first time that it is possible to test the hypotheses behind a minimal gene set using a chemically detailed, dynamic, whole-cell modeling approach. It has been demonstrated that it is possible to simulate a whole-cell whose behavior depends on its (i) metabolic rates and chemical state, (ii) genome in terms of expression of various genes, (iii) environment both in terms of direct nutrient starvation and competitive inhibition leading to starvation, and (iv) genomic sequence in terms of the locations of genes on the chromosome. All of these behaviors are exhibited by a single-cell model that makes reasonable assumptions about cellular biochemistry, reaction rates, gene expression, and the effect of discrete physiological events on the cell's behavior.

BIOGRAPHICAL SKETCH

Jordan Crawford Atlas was born in Boston, Massachusetts in September 1980 to Diane Crawford and Henry Atlas. He grew up with his mother Diane and younger brother Nathan. Jordan attended Memorial-Spaulding School, Charles E. Brown Junior High School, and Newton South High School in Newton, Massachusetts before moving to Brookline, Massachusetts and finally graduating from Brookline High School. In high school he developed a passion for long distance running, which stayed with him throughout college and graduate school, as well as an interest in computers, which guided most of his professional development. He was known by his friends for crying during touching scenes of video games and for spontaneously sitting up while asleep (such that he looked like a sleeping clamshell).

He attended the University of Massachusetts, Amherst and completed an undergraduate honors thesis with Dr. Susan Roberts in Chemical Engineering. As an undergraduate Jordan worked for two years as a Help Desk Consultant at the UMass Office of Information Technologies, and during summer and winter breaks he was an Information Technology Intern at Phase Forward, Inc. in Waltham, Massachusetts. As a junior, he became President of the UMass Shotokan Karate club. He graduated in 2003 with a Bachelor of Science in Chemical Engineering and a Bachelor of Science in Biochemistry.

In August 2003, Jordan began his studies in the School of Chemical and Biomolecular Engineering at Cornell University, and shortly thereafter joined the Shuler Research Group. His graduate work with Dr. Michael Shuler focused on computer modeling of bacterial cells. In 2005 he was awarded a Computational Science Graduate Fellowship (CSGF) and an Integrative Graduate Education and Research Traineeship (IGERT). As part of the CSGF

program, he completed a practicum in Theoretical Biology at Los Alamos National Laboratory in Los Alamos, New Mexico with Dr. James Faeder. He received the Edna O. and William C. Hooey Prize for Research in Chemical Engineering in November, 2010. His hobbies during graduate school included cross-country running, salsa, cooking, games, and puzzles.

In summer 2010, Jordan will start work at Microsoft in Redmond, Washington as a Software Developer in Test on the Enterprise Search Team.

To my family, to my friends, and to those who listen with empathy

ACKNOWLEDGEMENTS

I have always been better with faces than with names. As I admire the group portrait of all those who have helped make my graduate school experience possible, I worry that some of the names might get lost in the crowd. Nonetheless, I owe many people thanks and I will try to express that here.

My advisor, Michael Shuler, has made his support in many ways. Even when he seemed to be the most busy professor on campus, he always enthusiastically made time to talk about my work. He also patiently gave me advice on everything from teaching to career development. My committee members Matt DeLisa and David Wilson provided invaluable feedback at our periodic checkpoints, as did Jeff Varner after my defense. Teaching was an important part of my graduate education, and I am very grateful to have had the opportunity to develop as a teaching assistant with Michael Shuler, Abe Stroock, and Dave Putnam.

I owe thanks to several funding sources. My first year at Cornell was funded by a Cornell Graduate School Fellowship. Later, my work was funded by DOE grant DE-FG02-04ER63806 and by NYSTAR, the New York State Office of Science, Technology, and Academic Research. I also gratefully acknowledge support from the DOE Computational Science Graduate Fellowship Program (CSGF) of the Office of Science and National Nuclear Security Administration in the DOE under contract DE-FG02-97ER25308. The CSGF had a transformative influence on my professional development, and I am grateful for its four years of support. I would like to thank Martin Edelson, Rachel Huisman, and Jeanna Gingery for their guidance and assistance during my tenure with the fellowship. I am particularly grateful to James Faeder for serving as my fellowship practicum advisor at Los Alamos National Laboratory. He is one of

the most welcoming scientists I have ever met, and working with him for the summer was inspirational. I was also fortunate to participate in the Cornell Nonlinear Systems Integrative Graduate Education and Research Traineeship (IGERT), and I thoroughly enjoyed learning from John Guckenheimer and Steve Strogatz while I was in that program.

Assembling a dissertation was an unprecedented challenge for me. I appreciated the revisions and editing provided by Tricia Foley, Gretchen Mahler, and Anne Lin. Sue Payne went to great lengths to help me get the dissertation drafts to my committee on time. I am also thankful for support from other Cornell staff including Bonnie Sisco, Diana Guilford, and Shelby Clark-Shevalier.

I would like to thank Susan Roberts, my undergraduate research advisor at the University of Massachusetts, Amherst for encouraging me to pursue every subject that I was interested in and setting me on the path toward Cornell. When I first came to Cornell, Mariajose Castellanos introduced me to the Minimal Cell Model project and I appreciate the enthusiasm with which she sold the research. Evgeni Nikolaev, a colleague and mentor, showed me how to be mathematically rigorous and helped me discover what I was good at and what needed improvement. The Sethna Research Group, and in particular Ryan Gutenkunst, introduced me to the Python programming language, and their SloppyCell software package was integral to my work. Josh Waterfall, Chris Myers, and Robert Kuczenski also helped me learn how to use SloppyCell.

I have deep gratitude to my mother Diane for teaching me to love math and reading very early, and for going through extraordinary trouble to make sure that I grew up with the best education available in Massachusetts. Throughout my time in Ithaca she sent me her encouragement and positive thoughts and for

that I am grateful. I also appreciate that my brother Nathan always listened to my unsolicited advice and forgave me for the times I refused to give him any advice at all.

As I experienced the ups and downs of graduate life my friends kept me grounded and reminded me to take an occasional break. The Shuler Research Group was the best set of colleagues one could ask for. In particular, thanks to Tricia (who never got tired of my questions), Gretchen (who I followed here from Amherst), Mandy (who is a great listener), BJ (a kind office mate), and Rishard (who was willing to drive to Syracuse multiple times to watch anime with me). Thanks also to Peter (who can make me laugh so hard it causes pain), Dave (with whom I learned to tread water and poach eggs), Sara (for engaging conversations about how great Massachusetts is), Conor (who knows how to get me riled up), Sharon (who motivated me to keep running), Brian (who motivated me to keep gaming), Lydia (who had a special ability for seeking me out when I was in need of a friend), Eddie (for his unique perspective and friendship), Didi (a really cool person), Jeannie (for always being willing to hang out), Eric (who helped me pursue diverse interests), Geoff (a good friend and roommate), JoAnn (who helped me get to Ithaca), Jason (for being there for me when I went bad), and Anjali (who never realized that I learned more by answering her questions than she did from hearing my answers). Finally, I would like to thank my partner Anne Lin for her love and support, for making my last years in Ithaca the sweetest, and for being incredibly patient with me as I wrote this dissertation.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

GENOMICALLY DETAILED MODELS OF BACTERIAL CELLS

1.1 Introduction

"What is essential for life?" is one of the most fundamental questions we face. The complete reconstruction of a minimal cell *in silico* is key to fully understanding and identifying the underlying regulatory and organizational concepts central to life. Whole organism genome sequencing and high-throughput measurements provide opportunities for system-level analysis of whole organisms, or what has been termed "systems biology" (Ideker et al., 2001; Kitano, 2002). Systems biology investigates the behavior of all of the elements in a biological system while it is functioning (Ideker et al., 2001), which can help answer questions of essentiality for organisms. As a systems biology approach, the Minimal Cell Model (MCM) depicts the total functionality of a minimal cell and its explicit response to perturbations in the environment (Browning and Shuler, 2001).

A minimal cell is a hypothetical entity defined by the essential functions required for life (Castellanos et al., 2004). It is assumed that this cell exists in an environment with preformed nutrients, constant temperature, and constant pH. Although other research groups have the goal of experimentally constructing a minimal cell (Zimmer, 2003; Forster and Church, 2006, 2007; Lartigue et al., 2009), we seek to construct a dynamic model of such a cell. The model can be used as a tool to identify the organizing principles that relate the dynamic nonlinear functioning of the cell to the genome sequence.

The overall accomplishments of this research project build on a single-cell modeling approach pioneered in the late 1970s (Shuler and Dick, 1979; Domach, 1983; Shuler, 2005). The two main foci of this dissertation are (i) to develop more powerful and flexible computational techniques for analysis of coarse-grained bacterial cell models, and (ii) to develop a model of a hypothetical bacterium with the minimum number of genes necessary and sufficient to support sustained division, i.e. an MCM.

The long term impact of this work will make the MCM available to a wide audience. The model is available in the Systems Biology Markup Language (SBML) (Hucka et al., 2003, 2008) with model a simulator available in Python (Gutenkunst et al., 2007). Disseminating the model in this manner will provide practical guidance to researchers involved in bioprocesses, metabolic engineering, and interpretation of genomic information, especially in regard to techniques to construct "hybrid" models of real bacteria.

Sections 1.2-1.5 in Chapter 1 describe previous work done on the Cornell *E. coli* model, as well as the previous iterations of the MCM. The minimal gene set concept is introduced, and previous proposals for minimal gene sets are explained. A preview of the remainder of this dissertation is presented in Section 1.6.

1.2 Motivation

This research seeks to elucidate the common, essential features of a living cell (with a focus on chemoheterotrophic bacteria). In particular, a platform that allows investigators to unambiguously link genomic structure to cell

physiology is sought. A mathematical model of a "minimal cell" was constructed to provide a basis to better understand the design logic of cellular regulation (see Section 1.4 for a discussion of minimal cells). Although others have the goal of experimentally constructing a minimal cell (Zimmer, 2003; Forster and Church, 2006, 2007; Lartigue et al., 2009), this project aims to identify a minimal gene set and create a dynamic model of a bacterial cell that contains just those genes. Current estimates dictate that a minimal cell will have on the order of 200 to 300 genes and that all of these genes will have known functions. Most bacteria that exist in nature have on the order of 1,000 to 5,000 genes (e.g. *E. coli* has about 4,400 genes), and many of the products of these genes have unknown functions. Consequently a genomically detailed model of a real bacterium is neither practical (because it would be too large), nor desirable (because it would yield limited insight for the operation of genes with unknown functions). An MCM with a completely defined genome provides a platform to test, unambiguously, questions about how real whole cells must regulate themselves as well as a framework to model existing cells.

While the specific goal of this research is an MCM, the practical impact is broader. The minimal cell is a "learning model" used to probe the essence of a generalized cell response. The MCM and the techniques developed to produce such a model provide an essential foundation for "hybrid models" of bacterial cells. These models will use a "coarse-grained" overall model in which one embeds one or more genomically/molecularly detailed submodels (Shuler, 2005). The hybrid modeling strategy couples molecular details with a coarse-grained description of cellular processes and the extracellular environment. At the same time, this coupling can be linked to the chemical and genomic detail present in an MCM. All of the elements to form

a hybrid model (coarse-grained general structure, rapid estimation of kinetic parameters, and molecularly detailed modules of subsystems) are necessary to form the MCM. It provides a platform from which powerful mathematical techniques can be used both to determine criteria for robustness, and to rapidly prototype models of real cells.

Other broad impacts of this project include a greater insight into what is essential for life, which is a question of broad interest to both scientists and the lay public, as well as practical guidance to researchers involved in bioprocesses, metabolic engineering, and the interpretation of genomic information. Finally, there is a strong interest in using *in silico* models to connect bacterial genomic sequence information to physiological predictions. Specifically, biologists would like to be able to understand how changes to the genome sequence of an organism will affect its phenotypic behavior without necessarily making those genomic modifications *in vivo* or *in vitro*. Developing this understanding has practical implications in systems biology.

The MCM also has potential applications in synthetic biology. Foley and Shuler (2010) list five essential characteristics of a biological synthetic cell:

1. Robust mechanisms to control and correlate chromosome replication and cell division
2. Physically robust structure (e.g., cell envelope that allows high-density, large-scale culture without inducing cell lysis)
3. Decreased genetic drift (reduced mutation rates)
4. Simple and efficient transcription, translation, and regulatory systems to optimize flow of metabolic energy/resources to the design function

5. Mathematically defined interactions and predictable kinetics of the system

The fifth characteristic is the most important for this dissertation. A system with predictable kinetics would facilitate modeling, and having a chemically detailed model of a synthetic organism would allow an experimenter to test proposed modifications to the system and identify potential bottlenecks in production. The Shuler group has a long history of modeling bacterial cells to test modifications like these (Shuler and Dick, 1979; Domach and Shuler, 1984; Browning and Shuler, 2001; Castellanos et al., 2004; Atlas et al., 2008).

Another benefit of the proposed model is that it could lead to a better understanding of the behavior of real chemoheterotrophic bacteria, as well as more effective models of real bacteria. While an MCM suggests the essential components of regulation, deeper insight into the logic of cell regulation can also be achieved by introducing perturbations to the system where large changes can lead to failures in the model (i.e. cell death) and regulatory approaches could be found to counteract these changes (i.e. allow survival). As such, insight into cellular structure and regulation gained from the MCM become important for the metabolic engineering of cells and for the design of improved bioprocess strategies.

Finally, an MCM can be used as a platform to evaluate candidate minimal gene sets. There are several methods in popular use for estimating the core genes necessary for bacterial life, but there is currently no widely accepted method for testing the plausibility of those gene sets. Until synthetic biology offers a method to rapidly create a bacterium with a synthetic minimal genome on the lab bench, a simulation of a minimal cell is the best way to verify a particular gene set's viability.

1.3 Computer Models of Bacterial Cells

The MCM is built using a coarse-grained bacterial framework, which is one of several modeling strategies available to computational biologists and applied mathematicians studying whole bacterial cells. Using modeling, many investigators have made significant contributions to our understanding of bacterial metabolism. Some studies take advantage of detailed genomic information (Karp et al., 2004), while other models are based primarily on flux balance analysis, metabolic control theory, and mathematical techniques for optimization (Burgard et al., 2001; Burgard and Maranas, 2001; Edwards and Palsson, 2000; Edwards et al., 2002; Durot et al., 2009). These modeling techniques are all intrinsically static, and they have limited ability to predict aspects of cell regulation and dynamic response. Other investigators have proposed methods to directly incorporate dynamic (kinetic) information into models of central metabolism (Chassagnole et al., 2002). Moreover, while some have attempted to model whole cells (Tomita et al., 1999; Tomita, 2001), those models neglect important, non-metabolic aspects of cell growth (e.g. control of chromosome replication or cell division) because there is no formalism to handle such "events" in the context of a cell model.

Constraint-based models, including flux-balance analysis, have a large representation in the literature. Under the time scale of minutes, metabolite concentrations in cells are generally at steady levels and remain constant as long as environmental conditions do not change. Therefore, a modeler can use the law of conservation of mass to constrain the synthesis and consumption rates of those metabolites. This is expressed as a *stoichiometric* constraint based on the stoichiometric relation proposed by each reaction in the system under

study (Durot et al., 2009). For each metabolite, the mass balance constraint is written mathematically as $\sum s_j v_j = 0$, where s_j is the stoichiometric coefficient of the metabolite in reaction j, and v_j is reaction rate j. The stoichiometric constraints are supplemented with constraints regarding reaction reversibility and maximum reaction rate. The construction and applications of these models are reviewed in Durot et al. (2009), and there are several interesting applications available (Burgard and Maranas, 2001; Burgard et al., 2001; Edwards and Palsson, 2000; Edwards et al., 2002).

These studies, and many other similar ones, make important contributions toward our perception of systems biology. However, all of these approaches neglect the coupling between cell physiology and cell growth that is prevalent in physiological events such as chromosome replication. Descriptions that neglect this coupling may lead to conclusions that are inaccurate because they implicitly assume that the output of each pathway cannot influence any input into the same pathway (Schlosser and Bailey, 1990). Further, many of the models referenced above assume an objective function, which typically maximizes the growth rate. While such a function can be justified in the context of a specific short-term situation, the real objective function (e.g. survival of the organism) is more complex and involves issues such as the ability to grow robustly and in a variety of environmental conditions.

The Shuler group has previously developed a whole-cell model of *E. coli* that contains all of the functional elements for the cell to grow, divide, and respond to a wide variety of environmental perturbations. All chemical species are included, but lumped into pseudochemical groups. This "coarse-grained" model serves as the basis for our efforts to build an MCM. The Shuler group

first described a mathematical model of a single *E. coli* cell in 1979 (Shuler and Dick, 1979). While the *E. coli* model summarizes the physiological functionality required for a minimal cell, it does not capture explicitly the physical chemistry that supports those functions. It is unique in its natural coupling of metabolism, transport, and cellular events. At that time, it was the only model of an individual cell that did not dictate timing of cell division (e.g. growth rate) and cell size; instead, those aspects were outputs of the simulation. Also, it responded explicitly to concentrations of nutrients in the environment (Bailey, 1998). This base model (Domach, 1983) has been embellished with additional biological details to allow prediction of a wide-range of responses to environmental and genetic manipulations (Shuler, 1999). The initial model included only 18 pseudochemical species that represented large groups of related chemical species. Figure 1.1 lists the components of the *E. coli* model and graphically depicts their relationships.

The mathematical description of cellular functions that comprise the model is based on time-variant mass balances for each component. Each mass balance takes into account the component's synthesis (as a function of availability of precursors, energy, and relevant enzymes), utilization, and degradation. Stoichiometric coefficients for relating components through mass balances were derived primarily from published research, and in some cases, from experimental data. It is important to note that the model was not developed by using adjustable parameters to fit model predictions to experimental results, nor did the stoichiometric mass balances assume a steady-state (i.e. the amount of each component was allowed to vary with time). Despite the simplifications that were made in describing the cell, the model accurately predicts changes in cell composition, size, and shape, as well as the timing

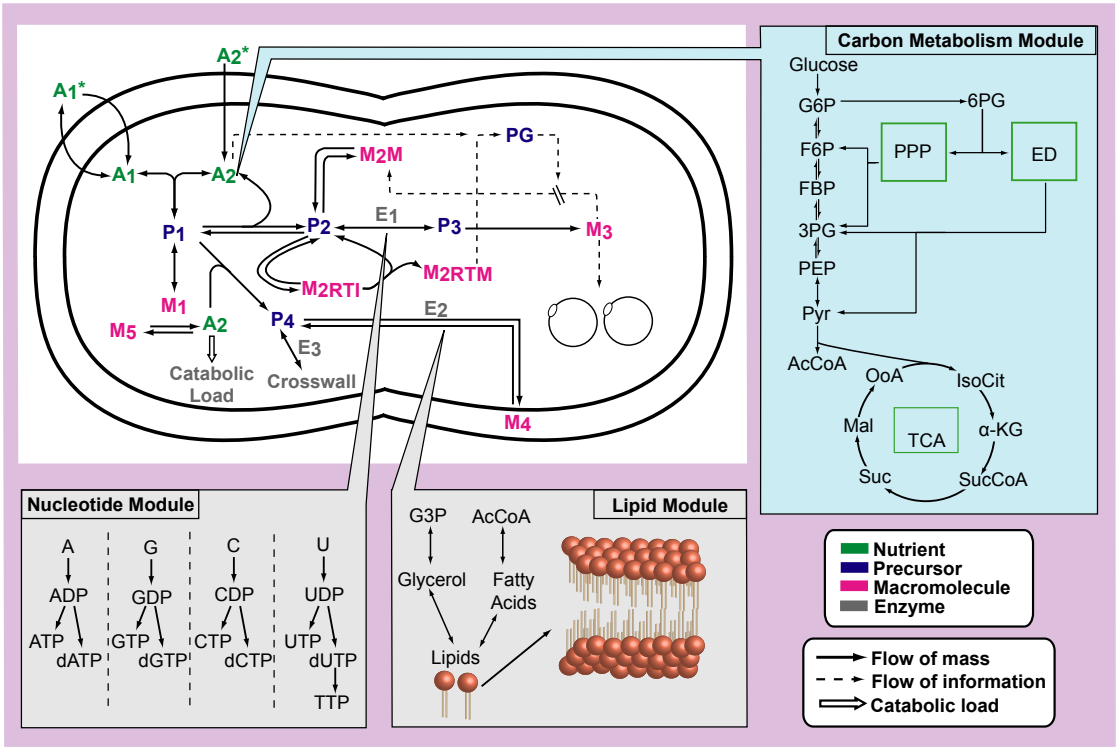


Figure 1.1: A schematic representation of the Single Cell Model and the modular approach to cell modeling. Grey boxes indicate chemically detailed modules that have been implemented, such as nucleotide metabolism (Castellanos et al., 2004) and lipid metabolism (Castellanos et al., 2007). The blue box illustrates an example of a potential new carbon metabolism module. Solid and dashed lines represent mass and information flows, respectively. 'Catabolic load' refers to glucose spent for energy metabolism, and 'Crosswall' refers to lipids spent for septum formation during cell division. Not all reactions and regulation information are depicted. PPP, ED, and TCA are the Pentose Phosphate Pathway, the Entner-Doudoroff Pathway, and the TCA Cycle. The labels in pathways represent lumped pseudo-species defined as: A_1 - ammonium ion, A_2 - glucose, P_1 - amino acids, P_2 - ribonucleotides, P_3 - deoxyribonucleotides, P_4 - membrane precursors, M_1 - protein, M_{2RTI} = immature stable RNA, M_{2RTM} - mature stable RNA, M_3 - DNA, M_4 - cell envelope, M_5 - glycogen, PG - ppGpp, E_1 - enzymes for conversion of P_2 to P_3, E_2 and E_3 - enzymes for cross-wall formation and cell envelope synthesis. * indicates species that are external to the cell (Domach et al., 1984).

of chromosome synthesis as a function of changes in external glucose and ammonium concentration (Domach et al., 1984; Lee et al., 1984; Shuler and Domach, 1983). The model also addresses important issues such as energy generation and the maintenance of the electropotential and chemical potential gradients across the cytosolic membrane by including a description of the cell's energy accounting process and the movement of H^+ ions (leaky protons) along the membrane (Shuler and Dick, 1979; Lee et al., 1984; Shuler and Domach, 1983). Two examples of stoichiometric mass balances for formation of precursors (amino acid) and macromolecules (RNA) are given in Equations 1.1 and 1.2.

$$\alpha_1 A_1 + \beta_2 A_2 + \ldots \rightarrow P_1 + \ldots \tag{1.1}$$

$$\gamma_2 P_2 \Rightarrow M_2 + \ldots \tag{1.2}$$

In Equations 1.1 and 1.2, α_1, β_1, and γ_2 are stoichiometric coefficients, and A_1, A_2, P_1, P_2, and M_2 are the masses of ammonium ion, glucose, amino acids, ribonucleotides, and total RNA, respectively. Chemical concentrations are measured in mass per cell, and stoichiometric balances are based on carbon and nitrogen. Equation 1.3 shows the corresponding requirements for phosphate energy coupled with the biosynthetic reactions.

$$\delta_{P1} ATP \rightarrow \delta_{P1}(ADP + Pi) \tag{1.3}$$

In Equation 1.3, δ_{P1} is a stoichiometric coefficient representing the average amount of ATP hydrolysis that must occur to supply the energy required for

synthesis of a specific amount of amino acids (P_1) per cell. Also the amount of reducing power formed and utilized is included in the accounting system.

The change in mass of a substance per cell per unit time can be found from a dynamic mass balance accounting for synthesis, import, export, and consumption. Note that this is not the same as concentration because the cell volume is changing. Equation 1.4 is an example mass balance for deoxyribonucleotides.

$$\frac{dP_3}{dt} = k_3 \cdot \left(\frac{K_{P3}}{K_{P3} + \frac{P_3}{V_C}}\right)\left(\frac{\frac{P_2}{V_C}}{K_{P3P2} + \frac{P_2}{V_C}}\right)\left(\frac{\frac{A_2}{V_C}}{K_{P3A2} + \frac{A_2}{V_C}}\right) \cdot E_1 - \gamma_3 \left(\frac{dM_3}{dt}\right) \quad (1.4)$$

where k_3 is the maximum rate of synthesis for deoxyribonucleotides formation ($time^{-1}$), K_{P3}, K_{P3P2}, and K_{P3A2} are saturation constants ($\frac{mass}{volume}$), γ_3 is a stoichiometric coefficient, and E_1 is the mass of enzyme E1 per cell (the rate limiting enzyme for conversion of ribonucleotides into deoxyribonucleotides). In Equation 1.4, the first term in brackets on the right hand side shows dependency based on deoxyribonucleotide concentration (P_3/V_C where V_C is cytosolic cell volume), the second term represents feedback inhibition of synthesis by ribonucleotide concentration (P_2/V_C), the third term indicates saturation-type dependence on glucose primarily for ability to supply energy (A_2/V_C), and the last term represents consumption to form DNA (M_3).

The original model explicitly describes discrete events that are typically ignored in other models (Nikolaev et al., 2006). For example, changes in gene dosage (the number of copies of a gene in a cell at a given time) depend on the replication fork position, and the completeness of cross-wall formation depends on the cell size and amount of cell membrane components synthesized. Other

biochemical details have been added in subsequent studies. For example, in one study, amino acids are differentiated into five families (Shu and Shuler, 1991) and the synthesis of ribosomes has been incorporated in greater detail (Laffend and Shuler, 1994a). These expansions allowed the study of the effects of amino acid supplementation (Shu and Shuler, 1991) and of competition between recombinant mRNA and ribosomal mRNA in the context of high translational activity (Laffend and Shuler, 1994a). The model was utilized extensively to improve the use of plasmids for recombinant protein production, e.g. (Laffend and Shuler, 1994a; Kim et al., 1987; Kim and Shuler, 1990, 1991; Laffend and Shuler, 1994b). The calculations have proved to be quite robust and results are reproducible. Bailey reviewed the importance of these contributions to the whole field of mathematical modeling in biochemical engineering (Bailey, 1998).

1.4 Minimal Cells

Before the current effort to construct an MCM is discussed, the minimal cell must be defined. The minimal cell concept can be traced back to the 1950s when Harold Morowitz and colleagues began to seek the smallest, autonomous, self-replicating entity (Morowitz, 1984). Because the genetic material of an organism defines its characteristics, what most succinctly defines a minimal cell is the makeup of its chromosome. Based on Morowitz's original concept, a minimal cell is defined as one possessing a minimal gene set, or a minimally sized list of genes that are both necessary and sufficient to promote sustained growth and division of a bacterial cell in some optimally supportive culture environment.

Various comparative genomic, genetic, and biochemical approaches have been used to estimate hypothetical minimal gene sets. Establishing a minimal gene set, or minimal gene sets, is an important step in synthetic biology. To prepare for incorporating a minimal gene set into an MCM, synthetic, natural, and experimental approaches to defining which genes belong in a minimal cell are considered. However, a reductionist approach that only considers each gene in the minimal gene set independently will be insufficient. It is necessary to evaluate how these cell systems functionally integrate (Moya et al., 2009).

1.4.1 Synthesis of Minimal Cells

One key focus of synthetic biology is the *de novo* construction of cells capable of performing important tasks like producing therapeutics or decontaminating waste streams (Foley and Shuler, 2010). There are bottom-up and top-down approaches to this goal. Bottom-up approaches attempt to synthesize a "living" cell that can reproduce, maintain homeostasis, and evolve without assuming the physiology of modern cells (Luisi et al., 2006). Alternatively, top-down approaches use modem cellular physiology as a starting point in the design of a synthetic cell (Forster and Church, 2006).

The J. Craig Venter Institute has been actively pursuing the goal of synthesizing a cell using a top-down approach. Toward this end, they successfully transplanted a complete *Mycoplasma mycoides* chromosome into a *Mycoplasma capricolum* cell which had its own genome removed (Lartigue et al., 2007). They also constructed a synthetic *Mycoplasma genitalium* genome *de novo* (Gibson et al., 2008). Finally, they took the entire genome from *M.*

mycoides, modified it in yeast using yeast genetic systems, and then transplanted the modified chromosome into *M. capricolum* (Lartigue et al., 2009). Together, these techniques put them very close to their ultimate goal of taking a wholly synthetic chromosome and using that as the starting genetic information for a new cell line. The only remaining steps are to clone a synthetic genome in yeast and then use that clone as the basis for a synthetic bacterium.

Although the Venter Institute is developing the technical procedures necessary for synthetic cell construction, another important step toward synthesizing a minimal cell is defining precisely what is in its genome. Furthermore, there are no examples of an experimental test of whether a proposed gene set is sufficient for driving cellular life. The first method used to consider which genes were both necessary and sufficient to drive life involved studying naturally occurring bacteria with minimized genomes (Morowitz, 1984).

1.4.2 Natural Examples of Minimized Gene Sets

There are some natural analogs of the hypothetical minimal cell that have evolutionarily reduced genome sizes. All known small-genome bacteria are associated with specialized lifestyles in stable environments, e.g., obligate symbiosis or specialized ecological niches (Moya et al., 2009). The two largest forces pushing a bacterial species toward genome reduction are symbiosis and resource economization, so it is not surprising that the smallest genomes in nature are all in prokaryotes living in symbiosis with other cells (Moya et al., 2009). Notable examples include: *Nanoarchaeum equitans*, a symbiotic

archaeon with a 490 kbp genome, or 536 protein-coding genes (Waters et al., 2003); *Buchnera aphidicola*, an endosymbiont of aphids with a 540 kbp genome, or 480 genes (Gil et al., 2002); and *Pachypsylla venusta*, an endosymbiont of hackberries with a 160 kbp genome, or 182 predicted ORFs (Nakabachi et al., 2006). Because it can be grown in pure cultures and has an extremely small genome size (580 kbp, 470 genes), *Mycoplasma genitalium* is considered the best living example of a minimal cell (Fraser et al., 1995); its genome represents a significant reduction from that of other well-studied bacteria such as *E. coli*, which has a 4,400 kbp genome. The *M. genitalium* genome developed through "top down" genomics, where genes are removed from an existing organism to provide a metabolically simpler cell (Maniloff, 1996). Thus, it exemplifies natural selection for a minimized genome.

As evidenced above, evolution (a "bottom up" approach) has suggested many forms of a minimal cell (Maniloff, 1996), but all of them can survive knockout experiments and are therefore not truly minimal. Estimates based on observation of naturally occurring bacteria suggest minimal gene sets in the range of 200-500 genes (Mushegian and Koonin, 1996; Hutchison et al., 1999; Koonin, 2000; Kobayashi et al., 2003; Gil et al., 2004; Glass et al., 2006). It has been proposed that a synthetic biology approach that takes advantage of enzymes with low substrate-specificity could drive the minimal gene set down to 100 or fewer genes (Murtas, 2007), but no minimal gene sets in that size range have been published.

1.4.3 Experimental Estimates of Minimal Gene Sets

There are genetic (Hutchison et al., 1999; Glass et al., 2006), comparative genomic (Tomita et al., 1999; Mushegian and Koonin, 1996; Koonin, 2000, 2003), and biochemical (Forster and Church, 2006; Luisi, 2002) approaches to establishing an *in vivo* minimal cell (Forster and Church, 2006). Taken together, these techniques go beyond naturally occurring minimization to propose minimal gene sets in the range of 200-400 genes.

Genetic approaches identify essential genes by large-scale gene disruption. Kobayashi et al. (2003) estimated 271 genes as the minimal gene set by systematically inactivating single genes in *Bacillus subtilis* using transposon mutagenesis experiments. Similar genetic methods have been used to estimate 1,490 essential genes in *Mycobacterium tuberculosis* (Lamichhane et al., 2003), 254 essential genes in *B. subtilis* (Itaya, 1995), and 382 essential genes in *M. genitalium* (Hutchison et al., 1999; Glass et al., 2006). Other efforts to determine gene essentiality using gene inactivation include (Forsyth et al., 2002) and (Gerdes et al., 2003). However, this experimental approach can lead to falsely labeling required genes as dispensable, which can derail any effort to create a minimal gene set (Forster and Church, 2006; Peterson and Fraser, 2001). Additionally a genetic approach can overestimate the minimal set substantially because genome scale knockouts could identify genes as essential even when the deletion only slows growth (Koonin, 2003).

In addition to estimates for a minimal gene set made using genetic techniques, estimates have been made using comparative genomics. Mushegian and Koonin estimated a set of about 250 genes as a minimal gene set after comparing the full genome sequences of *Haemophilus influenzae* and *M.*

genitalium (Mushegian and Koonin, 1996). In 2000, Koonin reviewed advances since their 1996 paper (Mushegian and Koonin, 1996) that demonstrate the complexity in using comparative genomics to establish a minimum gene set (Koonin, 2000). For example, of the 256 genes identified as essential in 1996, 15% were found to be dispensable in knockout experiments (Koonin, 2000). Many other computational analyses like these have been performed (Tomita et al., 1999; Nesb et al., 2001; Harris et al., 2003; Gil et al., 2003; Pál et al., 2006; Gabaldón et al., 2007; Carbone, 2006). However, comparative genomic approaches could yield either an over- or underestimation of minimal gene sets (Forster and Church, 2006). They are particularly prone to missing unrelated proteins with the same activity, or nonorthologous gene displacement (NOGD). Therefore, it is critical to develop a methodology for distinguishing among proposed minimal gene sets.

There have also been parallel efforts to determine the minimal set of cellular reactions or functions. Forster and Church described the main biochemical pathways that are necessary for essential bacterial functions, as well as an *in vitro* plan to synthesize a minimal cell (Forster and Church, 2006, 2007). They obtained a minimal genome with 151 genes for cellular information processing but omitted genes involved in major metabolic pathways (Forster and Church, 2006). Azuma and Ota (2009) determined the "minimal pathway maps", or the minimal set of autonomous pathways maps that could synthesize all required biomass components, for *E. coli* and *B. subtilis*. They found that pathways maps from the Kyoto Encyclopedia of Genes and Genomes (KEGG) were more likely to be conserved if they were involved in cellular information processing. This approach, while still computational, avoids the possibility of NOGD because a cellular function can be accepted into the minimal set regardless of NOGD.

The various approaches to determine a minimal gene set have been compiled and summarized in literature reviews (Gil et al., 2004; Forster and Church, 2006; Moya et al., 2009). Forster and Church (2006) conclude that the biochemical approach is still more promising than genetics or comparative genomics. They and others outline the steps necessary for synthesizing a minimal cell, primarily from genes found in *E. coli* (Zimmer, 2003; Forster and Church, 2006; Luisi, 2002). Forster lists the five gaps in our current knowledge that should be filled for the production of a synthetic minimal cell. The fourth among these is the lack of "biochemical parameters and computational models sufficiently detailed to predict the effects of alterations [in a near-minimal cell]" (Forster and Church, 2006). Similarly, Foley and Shuler (2010) list five essential characteristics of a biotechnological synthetic cell, the fifth being "mathematically defined interactions and predictable kinetics of (the) system". These claims illustrate the importance of the current work to produce a computational MCM.

In 2004, Gil et al. presented an enhanced review of all the previously proposed strategies for establishing a minimal gene set and proposed what they called the "core" of a minimal bacterial gene set (Gil et al., 2004). They started with a computational comparison of five sequenced endosymbionts: *Blochmannia floridanus*; *Wigglesworthia glossinidia*; and *Buchnera aphidicola*, strains BAp, BSg, and BBp (Gil et al., 2003). To that, they added in genes that had functional, but not sequence, similarity amongst the bacteria considered. They compared their gene set with the essential genes for *B. subtilis* (Kobayashi et al., 2003) and *E. coli* (Gerdes et al., 2003), as well as the computationally and experimentally derived minimal gene sets for *M. genitalium* (Mushegian and Koonin, 1996; Hutchison et al., 1999). Genes that were present in all five endosymbionts and that appeared to be essential in *Mycoplasmas* were

considered essential even if they were determined to be nonessential in bacteria with larger genomes (Gil et al., 2004). Finally, they analyzed the gene list to fill in gaps in metabolic pathways that are assumed to be essential. This resulted in a gene set with 206 protein coding genes (Gil et al., 2004). The total was later corrected to 207 protein coding genes to account for a step missing from the pentose phosphate pathway (Gabaldón et al., 2007).

The gene set proposed by Gil has the following features (Gil et al., 2004):

1. A virtually complete DNA replication machinery, composed of one nucleotide DNA binding protein, single-stranded binding protein (SSB), DNA helicase, primase, gyrase, polymerase III, and ligase.
2. A simple DNA repair system.
3. A virtually complete transcriptional machinery, including the three subunits of the RNA polymerase, a σ factor, an RNA helicase, and four transcriptional factors.
4. A nearly complete translational system.
5. Protein-processing, folding, secretion, and degradation.
6. Cell division driven by FtsZ only.
7. Two substrate transporters (PTS for glucose and PitA for inorganic phosphate).
8. ATP production via substrate-level phosphorylation.
9. Four enzymes from the non-oxidative branch of the pentose phosphate pathway.
10. Biosynthesis of phospatidylethanolamine from dihydroxyacetone phosphate and activated fatty acids.

11. Nucleotide biosynthesis from PRPP and free bases adenine, guanine, and uracil, which are obtained from the environment.
12. Cofactor biosynthesis from precursors obtained from the environment
13. No pathways for amino acid biosynthesis.
14. No protein transport systems for amino acids or inorganic ions (with the exception of phosphate).
15. No genes for stable RNA products (i.e. tRNA or rRNA), although they do define their proposed gene set as a minimal set of 'protein-coding' genes.

The implementation of these features in the MCM is discussed in Chapter 4. Gil et al. argue that there may be several possible minimal gene sets, saying "we should accept that there is no conceptual or experimental support for the existence of one particular form of minimal cell." In this work, one potential mechanism for distinguishing amongst minimal gene sets through computer modeling is presented.

1.5 Minimal Cell Model

Morowitz proposed that it should be possible to build a genomically complete computer model of a minimal cell (Morowitz, 1984). This dissertation considers construction of an MCM based on the gene set proposed by Gil et al. (2004). However, previous work to establish an MCM attempted to build a minimal gene set independently. In 2001, the Cornell *E. coli* model was first used by the Shuler group as a basis to construct an MCM that simulates a hypothetical bacterial cell with the minimum number of genes necessary to grow and divide

in an optimal environment (Browning and Shuler, 2001). The MCM has also been posed as a generalized model of chemoheterotrophic bacteria, which is called here the coarse-grained MCM. The original strategy for transitioning from the original Cornell single-cell model into the MCM was to sequentially replace 'pseudochemicals' and 'pseudoreactions' components of the model with distinct chemicals and detailed reactions (Castellanos et al., 2004, 2007). It is our belief that a detailed model of *E. coli* would not be computationally tractable because of its large number of gene products (Browning and Shuler, 2001). While it was not chemically detailed, the coarse-grained MCM was complete in terms of physiological function and was modular in its structure. A modular species is one that can be deconstructed into individual components while still maintaining the essential connectivity to other functions in the cell (Castellanos et al., 2004). Adding detail to different modules allows us to recombine those submodels into a functioning whole. This was the basic strategy for constructing a genomically and chemically detailed MCM.

The MCM is a functionally complete, system-level model formed by modification of a coarse-grained model of a single cell of *E. coli* (Browning and Shuler, 2001; Castellanos et al., 2004, 2007). The *E. coli* coarse-grained model can predict growth rate, cell composition, cell size and shape, response to addition to plasmids or specific genes, and genetic alterations as the nutrient environment is altered (Domach and Shuler, 1984; Kim and Shuler, 1990; Atlas et al., 2008). The coarse-grained model is based on lumped pseudo-chemical species. However, by "de-lumping" a pseudo-chemical species to provide genomic and chemical detail one can construct "modules" that can be incorporated into the overall model. The concept of modularity has been demonstrated by the inclusion of genomically/chemically detailed

nucleotide and lipid biosynthesis modules (Castellanos et al., 2004, 2007). Additionally, detailed genomic information about the location of DnaA binding boxes on the *E. coli* chromosome has been incorporated into the coarse-grained model to predict key features of DNA replication (Atlas et al., 2008). The MCM described here goes beyond these prior models to describe explicitly all genes in the cell, all chemical species, and incorporates mechanisms for most cellular processes.

The MCM focuses on essential functions while finding examples of gene products that can perform those functions. While the postulated set of minimal genes may change (e.g. if a new multifunctional protein is found), the set of essential functions is expected to stay relatively constant. Further, the technical difficulties associated with generating an experimental minimal cell and the ambiguities in interpretation of comparative genomic data promote the establishment of a theoretical computer model of a minimal cell. This model must be explicit about minimal functions and include a realistic set of proteins to accomplish these functions. This is, in essence, the primary objective of the proposed project and the most practical route to a minimal cell.

1.5.1 Previous Work on the Minimal Cell Model

The efficacy of constructing an MCM has been demonstrated in various proof of concept and validation studies (Browning and Shuler, 2001; Castellanos et al., 2004; Browning et al., 2004; Castellanos et al., 2007). To confirm that the concept of modularity was feasible within the modeling framework, the Shuler group added submodels for nucleotide and lipid metabolism to the MCM (Castellanos

et al., 2004, 2007). Both of these were selected as good starting points for the MCM because the pathways involved in the pseudoreactions for nucleotides and lipids involve a small number of genes (Mushegian and Koonin, 1996). The discussion below illustrates the principles of modularity by focusing on the development of the nucleotide module.

It has been demonstrated that it is not the exact values of parameters in the model that determine function, but that their values relative to one another is critical (Browning and Shuler, 2001). This hypothesis was tested by varying all kinetic rates by a scaling factor (or kinetic ratio), and it was found that growth rate scales directly with the kinetic ratio over about two orders of magnitude. At low values of growth rate, membrane energization becomes important and linearity is lost. Cell composition (e.g. protein/cell, RNA/cell, etc.) remains constant for a wide range of kinetic ratios. Further, relative growth rate changes for models with different kinetic ratios are essentially the same for a wide variety of perturbations to cell function (which also confirms the computational robustness of the model). The general physiological behavior of a variety of common bacteria (based on experiment) scales with a dimensionless growth rate. This suggests that the lessons from a hypothetical general cell model will be broadly applicable to chemoheterotrophic bacteria.

While the *M. genitalium* genome sequence suggests 25 genes can be associated with nucleotide metabolism and transport (Fraser et al., 1995), studies have estimated that as few as 10 of these may be essential (Hutchison et al., 1999; Mushegian and Koonin, 1996; Kobayashi et al., 2003). The pathway used in the coarse-grained MCM proposed by Castellanos et al. (2004) includes 11 functions (12 genes) and at the time it was published was the most efficient

(i.e., had the fewest genes) of any study with a complete pathway. An example of the equation used in the nucleotide model (Browning et al., 2004) describing the reduction of dUMP to synthesize dTMP by thymidylate synthase is shown in Equation 1.5, which is taken from Castellanos et al. (2004).

$$\frac{dP_{24dM}}{dt} = k_{12} \cdot K_{P24dMi} \cdot K_{P25dM} \cdot K_{P21T} \cdot V_C - \epsilon_9 \left(\frac{dP_{24dD}}{dt} \right) \tag{1.5}$$

Equation 1.5 makes use of the following three saturation term assignments for simplicity:

$$K_{P24dMi} = \left(\frac{K_{P24dM}}{K_{P24dM} + \frac{P_{24dM}}{V_C}} \right) \tag{1.6}$$

$$K_{P25dM} = \left(\frac{\frac{P_{25dM}}{V_C}}{K_{P24dM\text{-}P25dM} + \frac{P_{25dM}}{V_C}} \right) \tag{1.7}$$

$$K_{P21T} = \left(\frac{\frac{P_{21T}}{V_C}}{K_{P24dM\text{-}P21T} + \frac{P_{21T}}{V_C}} \right) \tag{1.8}$$

Above, k_{12} (time^{-1}) is the maximum rate of synthesis for dTMP synthesis; K_{P24dM}, $K_{P24dM\text{-}P25dM}$, $K_{P24dM\text{-}P21T}$ are saturation or equilibrium constants (mass/volume); P_{24dM}, P_{25dM}, P_{21T}, P_{24dD} are the mass per cell of dTMP, dUMP, ATP, and dUDP respectively, and V_C is the cell volume. All parameter values were estimated from experiments reported in the literature (Castellanos et al., 2004). A key demonstration in (Castellanos et al., 2004, 2007) is that a module can be de-lumped into genomically and chemically detailed components while maintaining a fully functional complete cell model. In essence, a coarse-grained minimal submodel gets embedded in the hybrid coarse-grained whole cell

model. Thus, we have established the concepts of modularity and connectivity and demonstrated that hybrid models of real bacteria are feasible.

Another important aspect of the original Cornell *E. coli* model was that it mechanistically coupled cell metabolism and growth with events such as chromosome replication and cell division (Shuler and Dick, 1979; Bailey, 1998). The original model for control of chromosome replication has been updated (Browning et al., 2004; Atlas et al., 2008) based on more recent experimental evidence (Hansen et al., 1991; Mahaffy and Zyskind, 1989; Donachie, 1993). While this model shares similarities with the initiator-titration model of Hansen et al. (1991), it includes ATP-bound DnaA as the active species rather than just DnaA. Both deterministic and stochastic versions of control of initiation of chromosome replication have been incorporated into the model. The stochastic version is necessary to determine robustness to intracellular fluctuations in concentrations.

1.5.2 Model Validation

Because the minimal cell is hypothetical, the MCM cannot be validated by a direct comparison to experimental data. However, the ability to predict the generalized behavior of chemoheterotrophic bacteria serves as a surrogate method to validate model predictions. The generalized behavior of such bacterium is used as a design performance constraint for model development. The advantage of doing this modeling exercise in a minimal cell is that every gene and gene product can be specified and the relationship of the system's dynamic response to perturbations can be explored; in real cells with genes

of unknown function there is always ambiguity in the interpretation of such an experiment. The MCM is built on the concept that all chemoheterotrophic microbes behave similarly. The model should demonstrate a general behavior that simulates how microbial growth responds to environmental changes. The model predictions have been compared to dimensionless microbial data (Browning and Shuler, 2001; Browning et al., 2004; Castellanos et al., 2004, 2007).

1.5.3 Current Challenges

The Shuler group has proposed the construction of an MCM as an alternative route to determine a minimal gene set for a chemoheterotrophic bacterial cell. An initial MCM has been constructed using the Cornell *E. coli* model as a basis and biological data (from several bacterial species) for development of new chemically detailed pathways. The generalized model has been compared to experimental data. While these approaches to adding realistic detail to the MCM do work, they are decidedly tedious.

For the current research, I have actively explored ways to increase the rate at which one can incorporate detail into the MCM by taking advantage of databases and new algorithm design. The completion of these tools and their application to the MCM is the focus of this dissertation. In contrast to previous work, the current research does not attempt to independently select which genes belong in the minimal gene set. Instead, the comprehensive minimal gene set proposed by Gil et al. (2004) is used as the basis for a new MCM.

There are two main applications of an MCM. It serves as:

1. A tool to test our understanding of biology.
2. A platform to test potential constructions of a real minimal cell (a.k.a. a synthetic cell), as well as to test minimal gene sets in general.

Successful construction of an experimental cell will require a system capable of replication and evolution fed by only small molecules (Forster and Church, 2006). Therefore, a successful MCM is defined in terms of its ability to simulate repeated replications in a nutrient rich environment comprised of small molecules provided in excess.

1.6 Preview of Subsequent Chapters

The Shuler group has pioneered the development of coarse-grained models of bacterial cells that incorporate chemical and genomic detail for systems of interest. These models are referred to as *hybrid* models. Chapter 2 presents a strategy for sensitivity analysis of hybrid models, with emphasis on the Cornell *E. coli* model.

In Chapter 3, an updated version of the Cornell *E. coli* model that incorporates a new deterministic model of the initiation of DNA replication controlled by the DnaA protein. This development was an important step toward developing a genomically detailed MCM because it was the first hybrid bacterial cell model to connect *detailed* genomic sequence information to the output of the simulation.

Chapter 4 describes the new Minimal Cell Model, including both the modeling structures used to create it as well as the submodels of metabolism

and physiological processes that drive it. The conventions and assumptions behind the MCM are presented, and the mathematical basis for the model is explained.

Chapter 5 presents some applications of the MCM. The MCM is used to calculate growth parameters for a minimal cell, as well as to predict the effects of various genetic and environmental manipulations

Finally, Chapter 6 describes the conclusions of this research and recommendations prompted from the new model.

A number of appendices have been included with supplementary information. These appendices are referred to throughout the dissertation, but of particular note are Appendix A, Model Naming Conventions, which explains the system used to name variables and parameters in the MCM, and Appendix E, which lists the full names of abbreviated chemical species in the MCM, as well as their initial masses in the cell.

REFERENCES

Atlas, J. C., Nikolaev, E. V., Browning, S. T., and Shuler, M. L. (2008). Incorporating genome-wide DNA sequence information into a dynamic whole-cell model of *Escherichia coli*: application to DNA replication. *IET Syst Biol*, 2(5), 369–382. doi:10.1049/iet-syb:20070079.

Azuma, Y. and Ota, M. (2009). An evaluation of minimal cellular functions to sustain a bacterial cell. *BMC Syst Biol*, 3, 111. doi:10.1186/1752-0509-3-111.

Bailey, J. E. (1998). Mathematical modeling and analysis in biochemical engineering: past accomplishments and future opportunities. *Biotechnology Progress*, 14(1), 8–20. doi:10.1021/bp9701269.

Browning, S. T., Castellanos, M., and Shuler, M. L. (2004). Robust control of initiation of prokaryotic chromosome replication: essential considerations for a minimal cell. *Biotechnology and Bioengineering*, 88(5), 575–584. doi:10.1002/bit.20223.

Browning, S. T. and Shuler, M. L. (2001). Towards the development of a minimal cell model by generalization of a model of *Escherichia coli*: Use of dimensionless rate parameters. *Biotechnology and Bioengineering*, 76(3), 187–192.

Burgard, A. P. and Maranas, C. D. (2001). Probing the performance limits of the *Escherichia coli* metabolic network subject to gene additions or deletions. *Biotechnology and Bioengineering*, 74(5), 364–375.

Burgard, A. P., Vaidyaraman, S., and Maranas, C. D. (2001). Minimal reaction sets for *Escherichia coli* metabolism under different growth requirements and uptake environments. *Biotechnology Progress*, 17(5), 791–797.

Carbone, A. (2006). Computational prediction of genomic functional cores specific to different microbes. *Journal of Molecular Evolution*, 63(6), 733–746. doi:10.1007/s00239-005-0250-9.

Castellanos, M., Kushiro, K., Lai, S. K., and Shuler, M. L. (2007). A genomically/chemically complete module for synthesis of lipid membrane in a minimal cell. *Biotechnology and Bioengineering*, 97(2), 397–409. doi:10.1002/bit.21251.

Castellanos, M., Wilson, D. B., and Shuler, M. L. (2004). A modular minimal cell model: purine and pyrimidine transport and metabolism. *Proceedings of the National Academy of Sciences of the United States of America*, 101(17), 6681–6686. doi:10.1073/pnas.0400962101.

Chassagnole, C., Noisommit-Rizzi, N., Schmid, J. W., Mauch, K., and Reuss, M. (2002). Dynamic modeling of the central carbon metabolism of *Escherichia coli*. *Biotechnology and Bioengineering*, 79(1), 53–73.

Domach, M. M. (1983). *Refinement and Use of a Structured Model of a Single Cell of Escherichia coli for the Description of Ammonia-Limited Growth and Asynchonous Population Dynamics*. Ph.D. thesis, Cornell University.

Domach, M. M., Leung, S. K., Cahn, R. E., Cocks, G. G., and Shuler, M. L. (1984). Computer model for glucose-limited growth of a single cell of *Escherichia coli* B/r-A. *Biotechnology and Bioengineering*, 26(9), 1140. doi:10.1002/bit.260260925.

Domach, M. M. and Shuler, M. L. (1984). A finite representation model for an asynchronous culture of *Escherichia coli*. *Biotechnology and Bioengineering*, 26(8), 877–884.

Donachie, W. D. (1993). The cell-cycle of *Escherichia coli*. *Annual Review of Microbiology*, 47, 199–230.

Durot, M., Bourguignon, P.-Y., and Schachter, V. (2009). Genome-scale models of bacterial metabolism: reconstruction and applications. *FEMS Microbiology Reviews*, 33(1), 164–190. doi:10.1111/j.1574-6976.2008.00146.x.

Edwards, J. S., Covert, M., and Palsson, B. (2002). Metabolic modelling of microbes: the flux-balance approach. *Environmental Microbiology*, 4(3), 133–140.

Edwards, J. S. and Palsson, B. O. (2000). The *Escherichia coli* MG1655 *in silico* metabolic genotype: Its definition, characteristics, and capabilities. *Proceedings of the National Academy of Sciences of the United States of America*, 97(10), 5528–5533.

Foley, P. L. and Shuler, M. L. (2010). Considerations for the design and construction of a synthetic platform cell for biotechnological applications. *Biotechnology and Bioengineering*, 105(1), 26–36. doi:10.1002/bit.22575.

Forster, A. C. and Church, G. M. (2006). Towards synthesis of a minimal cell. *Molecular Systems Biology*, 2, 45.

Forster, A. C. and Church, G. M. (2007). Synthetic biology projects *in vitro*. *Genome Research*, 17(1), 1–6. doi:10.1101/gr.5776007.

Forsyth, R. A., Haselbeck, R. J., Ohlsen, K. L., Yamamoto, R. T., Xu, H., et al. (2002). A genome-wide strategy for the identification of essential genes in *Staphylococcus aureus*. *Molecular Microbiology*, 43(6), 1387–1400.

Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., et al.

(1995). The minimal gene complement of *Mycoplasma genitalium*. *Science*, 270(5235), 397–403.

Gabaldón, T., Peretó, J., Montero, F., Gil, R., Latorre, A., et al. (2007). Structural analyses of a hypothetical minimal metabolism. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 362(1486), 1751–1762. doi:10.1098/rstb.2007.2067.

Gerdes, S. Y., Scholle, M. D., Campbell, J. W., Balazsi, G., Ravasz, E., et al. (2003). Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655. *Journal of Bacteriology*, 185(19), 5673–5684.

Gibson, D. G., Benders, G. A., Andrews-Pfannkoch, C., Denisova, E. A., Baden-Tillson, H., et al. (2008). Complete chemical synthesis, assembly, and cloning of a *Mycoplasma genitalium* genome. *Science*, 319(5867), 1215–1220. doi:10.1126/science.1151721.

Gil, R., Sabater-Munoz, B., Latorre, A., Silva, F. J., and Moya, A. (2002). Extreme genome reduction in *Buchnera* spp.: Toward the minimal genome needed for symbiotic life. *Proceedings of the National Academy of Sciences of the United States of America*, 99(7), 4454–4458.

Gil, R., Silva, F. J., Pereto, J., and Moya, A. (2004). Determination of the core of a minimal bacterial gene set. *Microbiology and Molecular Biology Reviews*, 68(3), 518–537.

Gil, R., Silva, F. J., Zientz, E., Delmotte, F., Gonzalez-Candelas, F., et al. (2003). The genome sequence of *Blochmannia floridanus*: comparative analysis of reduced genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 100(16), 9388–9393.

Glass, J. I., Assad-Garcia, N., Alperovich, N., Yooseph, S., Lewis, M. R., et al. (2006). Essential genes of a minimal bacterium. *Proceedings of the National Academy of Sciences of the United States of America*, 103(2), 425–430.

Gutenkunst, R. N., Atlas, J. C., Casey, F. P., Kuczenski, R. S., Waterfall, J. J., et al. (2007). SloppyCell, `http://sloppycell.sourceforge.net/`.

Hansen, F. G., Christensen, B. B., and Atlung, T. (1991). The initiator titration model - computer-simulation of chromosome and minichromosome control. *Research in Microbiology*, 142(2-3), 161–167.

Harris, J. K., Kelley, S. T., Spiegelman, G. B., and Pace, N. R. (2003). The genetic core of the universal ancestor. *Genome Research*, 13(3), 407–412. doi:10.1101/gr.652803.

Hucka, M., Finney, A., Sauro, H. M., Bolouri, H., Doyle, J. C., et al. (2003). The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4), 524–531.

Hucka, M., Hoops, S., Keating, S., Le Novre, N., Sahle, S., et al. (2008). Systems biology markup language (SBML) level 2: Structures and facilities for model definitions. *Nature Precedings*. doi:doi.org/10.1038/npre.2008.2715.1.

Hutchison, C. A., Peterson, S. N., Gill, S. R., Cline, R. T., White, O., et al. (1999). Global transposon mutagenesis and a minimal *Mycoplasma* genome. *Science*, 286(5447), 2165–2169.

Ideker, T., Galitski, T., and Hood, L. (2001). A new approach to decoding life: Systems biology. *Annual Review of Genomics and Human Genetics*, 2, 343–372.

Itaya, M. (1995). An estimation of minimal genome size required for life. *FEBS Letters*, 362(3), 257–260.

Karp, P. D., Arnaud, M., Collado-Vides, J., Ingraham, J., Paulsen, I. T., et al. (2004). The *E-coli* ecoCyc database: No longer just a metabolic pathway database. *Asm News*, 70(1), 25–30.

Kim, B. G., Good, T. A., Ataai, M. M., and Shuler, M. L. (1987). Growth-behavior and prediction of copy number and retention of ColE1-type plasmids in *Escherichia-coli* under slow growth-conditions. *Annals of the New York Academy of Sciences*, 506, 384–395.

Kim, B. G. and Shuler, M. L. (1990). A structured, segregated model for genetically modified *Escherichia coli* cells and its use for prediction of plasmid stability. *Biotechnology and Bioengineering*, 36(6), 581–592.

Kim, B. G. and Shuler, M. L. (1991). Kinetic-analysis of the effects of plasmid multimerization on segregational instability of ColE1 type plasmids in *Escherichia coli* B/R. *Biotechnology and Bioengineering*, 37(11), 1076–1086.

Kitano, H. (2002). Systems biology: A brief overview. *Science*, 295(5560), 1662–1664.

Kobayashi, K., Ehrlich, S. D., Albertini, A., Amati, G., Andersen, K. K., et al. (2003). Essential *Bacillus subtilis* genes. *Proceedings of the National Academy of Sciences of the United States of America*, 100(8), 4678–4683.

Koonin, E. V. (2000). How many genes can make a cell: The minimal-gene-set concept. *Annual Review of Genomics and Human Genetics*, 1, 99–116.

Koonin, E. V. (2003). Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nature Reviews Microbiology*, 1(2), 127–136.

Laffend, L. and Shuler, M. L. (1994a). Ribosomal-protein limitations in

Escherichia coli under conditions of high translational activity. *Biotechnology and Bioengineering*, 43(5), 388–398.

Laffend, L. and Shuler, M. L. (1994b). Structured model of genetic-control via the *lac* promoter in *Escherichia coli*. *Biotechnology and Bioengineering*, 43(5), 399–410.

Lamichhane, G., Zignol, M., Blades, N. J., Geiman, D. E., Dougherty, A., et al. (2003). A postgenomic method for predicting essential genes at subsaturation levels of mutagenesis: Application to *Mycobacterium tuberculosis*. *Proceedings of the National Academy of Sciences of the United States of America*, 100(12), 7213–7218.

Lartigue, C., Glass, J. I., Alperovich, N., Pieper, R., Parmar, P. P., et al. (2007). Genome transplantation in bacteria: changing one species to another. *Science*, 317(5838), 632–638. doi:10.1126/science.1144622.

Lartigue, C., Vashee, S., Algire, M. A., Chuang, R.-Y., Benders, G. A., et al. (2009). Creating bacterial strains from genomes that have been cloned and engineered in yeast. *Science*, 325(5948), 1693–1696. doi:10.1126/science.1173759.

Lee, A. L., Ataai, M. M., and Shuler, M. L. (1984). Double-substrate-limited growth of *Escherichia coli*. *Biotechnology and Bioengineering*, 26(11), 1398–1401.

Luisi, P. L. (2002). Toward the engineering of minimal living cells. *Anatomical Record*, 268(3), 208–214.

Luisi, P. L., Ferri, F., and Stano, P. (2006). Approaches to semi-synthetic minimal cells: a review. *Naturwissenschaften*, 93(1), 1–13. doi:10.1007/s00114-005-0056-z.

Mahaffy, J. M. and Zyskind, J. W. (1989). A model for the initiation of replication in *Escherichia coli*. *Journal of Theoretical Biology*, 140(4), 453–477.

Maniloff, J. (1996). The minimal cell genome: "on being the right size". *Proceedings of the National Academy of Sciences of the United States of America*, 93(19), 10004–10006.

Morowitz, H. J. (1984). The completeness of molecular-biology. *Israel Journal of Medical Sciences*, 20(9), 750–753.

Moya, A., Gil, R., Latorre, A., Peret, J., Garcilln-Barcia, M. P., et al. (2009). Toward minimal bacterial cells: evolution vs. design. *FEMS Microbiology Reviews*, 33(1), 225–235. doi:10.1111/j.1574-6976.2008.00151.x.

Murtas, G. (2007). Question 7: construction of a semi-synthetic minimal cell: a model for early living cells. *Origins of Life and Evolution of the Biosphere*, 37(4-5), 419–422. doi:10.1007/s11084-007-9090-5.

Mushegian, A. R. and Koonin, E. V. (1996). A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 93(19), 10268–10273.

Nakabachi, A., Yamashita, A., Toh, H., Ishikawa, H., Dunbar, H. E., et al. (2006). The 160-kilobase genome of the bacterial endosymbiont *Carsonella*. *Science*, 314(5797), 267. doi:10.1126/science.1134196.

Nesb, C. L., Boucher, Y., and Doolittle, W. F. (2001). Defining the core of nontransferable prokaryotic genes: the euryarchaeal core. *Journal of Molecular Evolution*, 53(4-5), 340–350. doi:10.1007/s002390010224.

Nikolaev, E., Atlas, J., and Shuler, M. L. (2006). Computer models of bacterial

cells: from generalized coarse-grained to genome-specific modular models. *Journal of Physics: Conference Series*, 46, 322–326.

Pál, C., Papp, B., Lercher, M. J., Csermely, P., Oliver, S. G., et al. (2006). Chance and necessity in the evolution of minimal metabolic networks. *Nature*, 440(7084), 667–670. doi:10.1038/nature04568.

Peterson, S. N. and Fraser, C. M. (2001). The complexity of simplicity. *Genome Biology*, 2(2), 1–8.

Schlosser, P. M. and Bailey, J. E. (1990). An integrated modeling-experimental strategy for the analysis of metabolic pathways. *Mathematical Biosciences*, 100(1), 87–114.

Shu, J. and Shuler, M. L. (1991). Prediction of effects of amino-acid supplementation on growth of *Escherichia coli* B/r. *Biotechnology and Bioengineering*, 37(8), 708–715.

Shuler, M. L. (1999). Single-cell models: promise and limitations. *Journal of Biotechnology*, 71(1-3), 225–228.

Shuler, M. L. (2005). Computer models of bacterial cells to integrate genomic detail with cell physiology. *Proceedings of the KBM International Symposium on Microorganisms and Human Well-Being, June 30-July 2005, Seoul Korea*.

Shuler, M. L. and Dick, C. (1979). A mathematical model for the growth of a single bacterial cell. *Annals of the New York Academy of the Sciences*, 326, 35–55.

Shuler, M. L. and Domach, M. M. (1983). Mathematical-models of the growth of individual cells - tools for testing biochemical-mechanisms. *ACS Symposium Series*, 207, 93–133.

Tomita, M. (2001). Whole-cell simulation: a grand challenge of the 21st century. *Trends in Biotechnology*, 19(6), 205–210.

Tomita, M., Hashimoto, K., Takahashi, K., Shimizu, T. S., Matsuzaki, Y., et al. (1999). E-CELL: software environment for whole-cell simulation. *Bioinformatics*, 15(1), 72–84.

Waters, E., Hohn, M. J., Ahel, I., Graham, D. E., Adams, M. D., et al. (2003). The genome of nanoarchaeum equitans: Insights into early archaeal evolution and derived parasitism. *Proceedings of the National Academy of Sciences of the United States of America*, 100(22), 12984–12988.

Zimmer, C. (2003). Genomics - Tinker, tailor: Can Venter stitch together a genome from scratch? *Science*, 299(5609), 1006–1007.

CHAPTER 2

MATHEMATICAL ANALYSIS OF A SINGLE CELL MODEL OF *ESCHERICHIA COLI*

2.1 Abstract

We discuss a modular modeling framework to rapidly develop mathematical models of bacterial cells that would explicitly link genomic details to cell physiology and population response. An initial step in this approach is the development of a coarse-grained model, describing pseudo-chemical interactions between lumped species. A hybrid model of interest can then be constructed by embedding genome-specific detail for a particular cellular subsystem (e.g. central metabolism), called here a module, into the coarse-grained model. Specifically, a new strategy for sensitivity analysis of the cell division limit cycle is introduced to identify which pseudo-molecular processes should be delumped to implement a particular biological function in a growing cell (e.g. ethanol overproduction or pathogen viability). To illustrate the modeling principles and highlight computational challenges, the Cornell coarse-grained model of *Escherichia coli* B/r-A is used to benchmark the proposed framework.

A general sensitivity and control analysis of periodically forced reaction

[1]Nikolaev, E.V., Atlas, J.C., and Shuler, M.L., 2006, "Computer models of bacterial cells: from generalized coarse-grained to genome-specific modular models", *Journal of Physics: Conference Series*, vol. 46, pp. 322-326, ©2006 IOP Publishing Ltd., `http://iopscience.iop.org/1742-6596`

networks with respect to small perturbations in arbitrary network's parameters and forcing frequency was also published (Nikolaev, Atlas, and Shuler, 2007). The abstract to this work is presented in Appendix G.

2.2 Introduction

Microbial genome sequences have become a central bioinformatic resource in modern biology by providing access to thousands of accurate metabolic reconstructions of completely annotated genomes (Overbeek et al., 2005), as well as genome-scale reaction networks and detailed stoichiometric models (Palsson, 2004). Despite their dominance and fundamental importance, intrinsically static metabolic reconstructions and stoichiometric models are, by themselves, insufficient to explicitly relate genomes to dynamic physiologic responses. The predictive capability of stoichiometric models is limited to the calculation of instant phenotype snapshots under fixed medium conditions. Therefore, such models cannot capture dynamic changes in metabolite concentrations, protein machinery, cell geometry, etc. At the same time, dynamic models are subject to difficulties in terms of sensitivity, stability, and robustness.

We developed a modeling approach to relate genomic data to dynamic intracellular processes: generalized hybrid models (Shuler, 2005). Hybrid models start with a functionally complete coarse-grained model which explicitly links DNA replication, metabolism, cell division, and geometry to the external environment (Domach et al., 1984). Such models can describe changes in energy and redox equivalents, RNA transcripts, transport, etc. The

availability of detailed metabolic reconstructions (Overbeek et al., 2007) and genome-scale reaction networks (Palsson, 2004) can significantly accelerate the development of genome-specific modules (Shuler, 2005) which can then be reused in many large-scale computer hybrid models for a variety of completely annotated genomes.

Another advantage of generalized models is that they combine a detailed summary of the functionality required to sustain the cell's life with modest-size model's complexity. Such models are thus an ideal platform for the development of computationally tractable systems biology concepts and biomathematics approaches. In this chapter, mathematical and computational approaches to evaluate the model's sensitivity and robustness are discussed. Specifically, a new strategy for the extension of Metabolic Control Analysis (MCA) (Heinrich and Schuster, 1996) to limit cycles is introduced to identify which pseudomolecular processes should be delumped to implement a particular biological function in hybrid cell models. To illustrate the modeling principles and highlight computational issues, the updated Cornell *Escherichia coli* B/r-A model is used to benchmark the framework.

2.3 The Model and Computational Frameworks

The Cornell *Escherichia coli* model, depicted in Figure 2.1, represents a single cell of *E. coli* growing in a glucose-ammonium medium. The model describes metabolism, DNA replication, and cell geometry (Domach et al., 1984). The modeling principles include: (i) the aggregation of cellular compounds into a manageable number of lumped species, (ii) the use of pseudochemical

reactions and accurate stoichiometry, and (iii) the evaluation of as many kinetic parameters as possible from independent measurements. A recently updated model includes 36 ODEs for metabolism and DNA replication, one algebraic approximation of the ribosomal protein biosynthesis, one algebraic equation to monitor the septum growth, and 31 discrete events describing instant changes in the model's parameters and state variables (e.g. changes in gene dosage, cell division, etc.). Dynamic systems describing a smooth evolution coupled with discrete transitions are called *hybrid*. The *E. coli* model is thus a *hybrid differential-algebraic equation* (HDAE), implemented in MATLAB®, C++, and Systems Biology Markup Language (SBML).

The *E. coli* model event network is depicted in Figure 2.2, where nodes correspond to events and arrows indicate how events can cause one another. We find that DNA replication initiation is the most connected node (i.e. E3) signifying its central role in the cell cycle. Although MATLAB® event detection is used, additional means are needed to identify "secondary" events induced by changes in the HDAE definition at each event. To catch all events, a general event detection algorithm has been developed.

A stationary cell division cycle corresponds to a periodic solution of the HDAE. One way to study periodic solutions is to compute a first return or Poincaré map $\mathbf{P}(\mathbf{s}, \mathbf{p})$, relating any two cell states over least period T, $\mathbf{s_{t+T}} = \mathbf{P}(\mathbf{s_t}, \mathbf{p})$. Here vector $\mathbf{s}$ includes masses, concentrations and numbers of molecules, and $\mathbf{p}$ includes model parameters (i.e. kinetic rates). Each event l is defined by the zero level of a scalar function $F_l(\mathbf{s}, \mathbf{p})$, $F_l(\mathbf{s}, \mathbf{p}) = \mathbf{0}$. We assume that $\mathbf{s_t}$ never corresponds to any event, i.e. $F_l(\mathbf{s_t}, \mathbf{p}) \neq \mathbf{0}$ for all $l = 1, \ldots, L$, where L is the total number of events. Let $\mathbf{H_l}$ be an event transition,

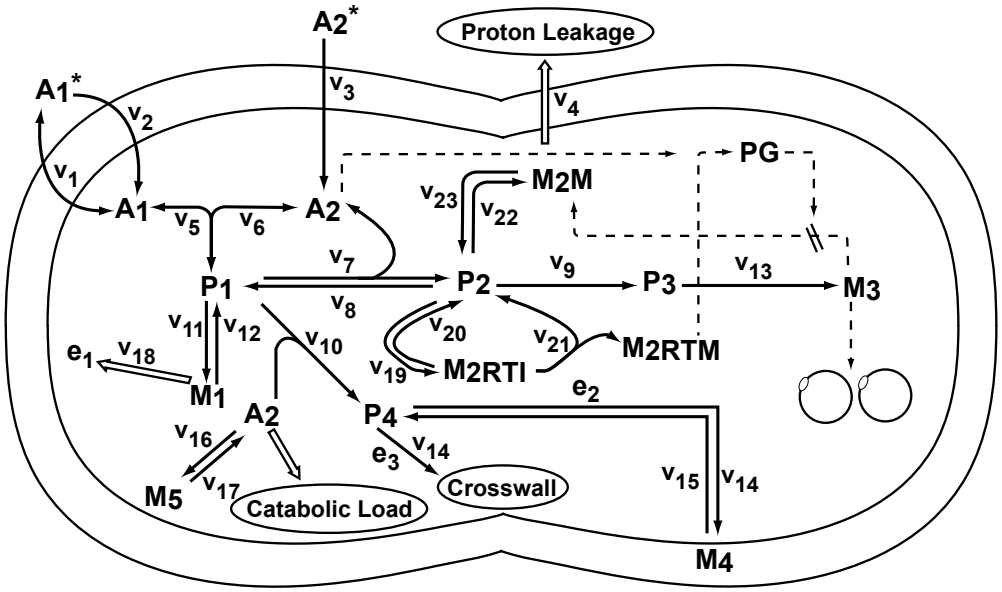


Figure 2.1: The Cornell coarse-grained *E. coli* model, which includes A_1 = ammonium ion, A_2 = glucose, P_1 = amino acids, P_2 = ribonucleotides, P_3 = deoxyribonucleotides, P_4 = cell envelope precursors, M_1 = proteins, M_{2RTI} = immature 'stable' RNA, M_{2RTM} = mature stable RNA, M_{2M} = messenger RNA, M_3 = DNA, M_4 = the nonprotein part of cell envelope, M_5 = glycogen, PG = ppGpp, e_1 = enzyme in the conversion of P2 to P3, and e_2 and e_3 = enzymes for cell envelope and cross-wall formation. Solid lines indicate flow of material, while dashed lines indicate flow of information. Here v_i is the rate of pseudoreaction i.

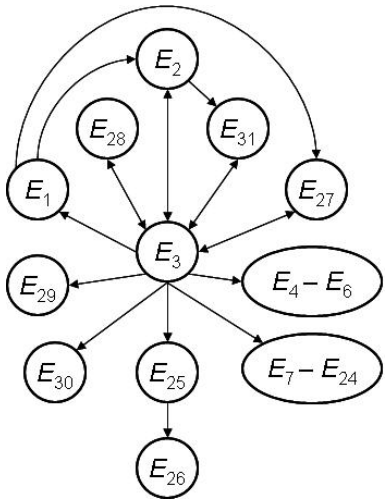


Figure 2.2: Events: E_1 = completion of DNA methylation, E_2 = transition of replicon state, E_3 = DNA replication initiation, E_4E_6 = changes in dnaA dosage, E_7 E_{24} = changes in *rrn*-operon dosage, E_{25} = DNA replication termination, E_{26} = cell division, E_{27} = the ability of DnaA-ATP and E_{28} = the ability of DnaA-ATP to bind high affinity DNA boxes, E_{29} = the ability of DnaA to bind medium affinity boxes, E_{30} = the ability of DnaA to bind nonspecific boxes, and E_{31} = the ability of DnaA to bind the triggering R5 box in *oriC*.

$(\mathbf{s}^+, \mathbf{p}^+) = \mathbf{H_l}(\mathbf{s}^-, \mathbf{p}^-)$, where $(\mathbf{s}^+, \mathbf{p}^+)$ and $(\mathbf{s}^-, \mathbf{p}^-)$ are chosen right after and just prior event l, respectively. If all DAEs defined between events, $\mathbf{H_l}(\mathbf{s}, \mathbf{p})$, and $F_l(\mathbf{s}, \mathbf{p})$ are smooth, then $\mathbf{P}(\mathbf{s}, \mathbf{p})$ is smooth in both $\mathbf{s}$ and $\mathbf{p}$. This follows from the decomposition of $\mathbf{P}(\mathbf{s}, \mathbf{p})$ into superposition of the smooth time shifts $\mathbf{Q_k}$ along the trajectories of the corresponding DAEs and the event transitions (Equation 2.1).

$$\mathbf{P}(\mathbf{s_{t+T}}, \mathbf{p}) = \mathbf{Q_{L+1}} \circ \mathbf{H_L} \circ \mathbf{Q_{L-1}} \circ \ldots \mathbf{Q_1} \circ \mathbf{H_1} \circ \mathbf{Q_0}(\mathbf{s_t}, \mathbf{p}) \tag{2.1}$$

Here, $\mathbf{Q_0}(\mathbf{s_t}, \mathbf{p})$ is the transition between $\mathbf{s_t}$ and the first event, and $\mathbf{Q_{L+1}}(\mathbf{s_{t+T}}, \mathbf{p})$ is the transition between the last event and $\mathbf{s_{t+T}}$. A fixed point

$\mathbf{s_0}$ of $\mathbf{P}(\mathbf{s}, \mathbf{p})$ uniquely defines the stationary cell division cycle. The fixed point $\mathbf{s_0}$ can be found from the nonlinear equation $\mathbf{s_0} = \mathbf{P}(\mathbf{s_0}, \mathbf{p})$ using Newton-like solvers.

An alternative approach to calculate periodic solutions is to solve a periodic multi-point boundary value problem (BVP) for a discrete closed orbit $\mathbf{z} = [(\mathbf{t_0}, \mathbf{s_0}, \mathbf{p_0}), \ldots, (\mathbf{t_{L+1}}, \mathbf{s_{L+1}}, \mathbf{p_{L+1}})], \mathbf{s_0} = \mathbf{s_{L+1}}$ (Phipps, 2003; Doedel et al., 2004). Here each $(\mathbf{s_l}, \mathbf{p_l})$ is chosen just prior to event l. The unknown period T, event times $t_1, \ldots, t_L$, and states $\mathbf{s_l}$ can be found using Newton solvers. Additional mesh points between events and a phase condition are needed to increase the accuracy and uniquely determine the periodic solution, respectively (Doedel et al., 2004).

Typical time courses are shown in Figures 2.3 and 2.4. We find that while changes in some species (e.g. ammonium ions, proteins) look "smooth" between divisions, other species (e.g. different forms of DnaA molecules or RNA transcripts) experience complex behavior throughout the entire cell cycle.

2.4 Sensitivity and Stability of the Cell Division Cycle

Sensitivity analysis is an important tool for model evaluation as well as for quantifying effects of parameter values on model predictions (Tomović and Vukobratovitć, 1972). Specifically, it is important to characterize the relative significance of various intracellular dynamic processes for modeling a growing cell. This can be done by an appropriate extension of MCA to the case of self-oscillations in autonomous hybrid systems. Let $\mathbf{s}(\mathbf{t}, \mathbf{p})$ be a stable periodic solution with least period $T(\mathbf{p})$, where $\mathbf{p}$ is a vector of the system's parameters.

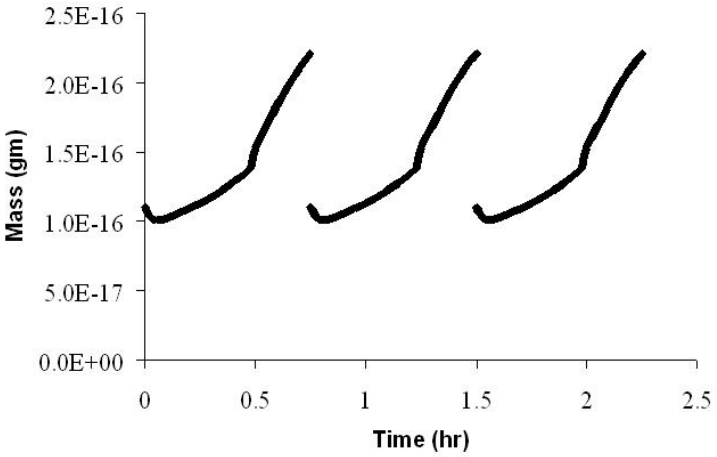


Figure 2.3: The mass of ammonium ions A_1 in the *E. coli* model.

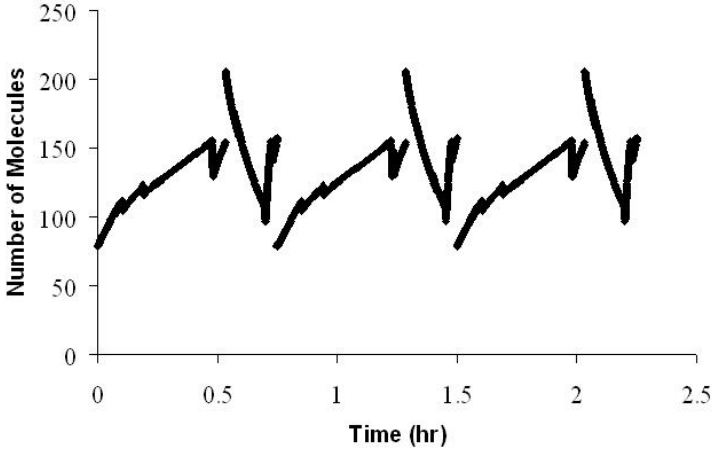


Figure 2.4: Free $DnaA\text{-}ATP$ in the *E. coli* model.

A well known property of such solutions in smooth dynamic systems is that a first-order sensitivity function (Equation 2.2),

$$\mathbf{u_k}(\mathbf{t},\mathbf{p}) = \frac{\partial \mathbf{s}(\mathbf{t},\mathbf{p})}{\partial \mathbf{p_k}} \tag{2.2}$$

is generally unbounded when time tends to infinity (Tomović and Vukobratovitć, 1972; Kholodenko et al., 1997). Here $k = 1,\ldots,K$, where K is the number of parameters. The key idea to understand and overcome this analytic difficultly can be seen from the differentiation of periodic condition $\mathbf{s}(\mathbf{t_m}(\mathbf{p}),\mathbf{p}) = \mathbf{s}(\mathbf{t_0},\mathbf{p})$ with respect to any scalar parameter p_k (Equation 2.4).

$$\begin{aligned} \mathbf{u_k}(\mathbf{t_m}(\mathbf{p}),\mathbf{p}) &= \mathbf{u_k}(\mathbf{t_0},\mathbf{p}) - \frac{(\mathbf{t_m}(\mathbf{p}-\mathbf{t_0})}{\mathbf{p_k}} \cdot \frac{\partial \ln \mathbf{T}(\mathbf{p})}{\partial \ln \mathbf{p_k}} \cdot \frac{\mathbf{ds}(\mathbf{t_0},\mathbf{p})}{\mathbf{dt}}, \\ t_m(\mathbf{p}) &= t_0 + mT(\mathbf{p}), \\ m &= 1,2,\ldots \end{aligned} \tag{2.3}$$

We find that $\mathbf{u_k}(\mathbf{t_m}(\mathbf{p}),\mathbf{p})$ becomes unbounded as the number of cell cycles m infinitely increases. By rescaling time as $\tau = 2\pi t/T(\mathbf{p})$, the unbounded $T(\mathbf{p})$-dependent term can be eliminated from Equation 2.4,

$$\begin{aligned} \mathbf{U_k}(\tau + \mathbf{2}\pi,\mathbf{p}) &= \mathbf{U_k}(\tau,\mathbf{p}), \\ \mathbf{U_k}(\tau,\mathbf{p}) &= \frac{\partial \mathbf{S}(\tau,\mathbf{p})}{\partial p_k}, \\ \mathbf{S}(\tau,\mathbf{p}) &= s\left(\frac{T(\mathbf{p})\tau}{2\pi},\mathbf{p}\right) \end{aligned} \tag{2.4}$$

The normalized period (i.e. $T = 2\pi$) and frequency (i.e. $\omega = 1$) are now independent of any parameter. Dimensionless time τ can be interpreted as

the phase ϕ of the cell cycle, $\phi = \tau(\mathrm{mod}\, 2\pi)$. Similar time scaling is used in the sensitivity theory (Tomović and Vukobratovitć, 1972) and the bifurcation analysis of limit cycles (Doedel et al., 2004). Using Equation 2.5, the summation laws quantifying the ability of enzymes to influence periodic processes can be readily obtained for cellular systems where the enzyme activities enter reaction rates (Equation 2.5) linearly (Heinrich and Schuster, 1996; Kholodenko et al., 1997).

$$v_j(\mathbf{s}, \mathbf{p_j}) = \mathbf{p_j} \mathbf{w_j}(\mathbf{s}) \tag{2.5}$$

Here p_j and $w_j(\mathbf{s})$ are the catalytic activity the turnover rate of enzyme j, respectively. Indeed, rescaling all p_j by the same nonzero factor λ will merely result in the change of the time scale (Heinrich and Schuster, 1996; Kholodenko et al., 1997).

$$\begin{aligned} T(\lambda \mathbf{p}) &= \frac{T(\mathbf{p})}{\lambda}, \\ s_i(t, \lambda \mathbf{p}) &= s_i(\lambda t, \mathbf{p}), \\ v_j(t, \lambda \mathbf{p}) &= \lambda v_j(\lambda t, \mathbf{p}) \end{aligned} \tag{2.6}$$

Here $v_j(t, \mathbf{p}) = \mathbf{v_j}(\mathbf{s}(\mathbf{t}, \mathbf{p}), \mathbf{p})$. Using Equations 2.5, 2.7 and flux notation $J_j(\tau, \lambda \mathbf{p}) = \mathbf{v_j}(\mathbf{S}(\tau, \mathbf{p}), \mathbf{p})$, we obtain Equation 2.8.

$$\begin{aligned} T(\lambda \mathbf{p}) &= \frac{T(\mathbf{p})}{\lambda}, \\ S_i(\tau, \lambda \mathbf{p}) &= S_i(\tau, \mathbf{p}), \\ J_j(\tau, \lambda \mathbf{p}) &= \lambda J_j(\tau, \mathbf{p}) \end{aligned} \tag{2.7}$$

The summation laws limiting changes in the period and shape of the limit cycle parameterized in Equation 2.5 follow from the differentiation of identities (Equation 2.8) with respect to non-zero scaling factor λ at $\lambda = 1$.

$$\begin{aligned} \sum_{j=1}^{K} \frac{\partial \ln T(\mathbf{p})}{\partial \ln p_j} &= -1, \\ \sum_{j=1}^{K} \frac{\partial \ln S_i(\tau, \mathbf{p})}{\partial \ln p_j} &= 0, \\ \sum_{j=1}^{K} \frac{\partial \ln J_i(\tau, \mathbf{p})}{\partial \ln p_j} &= 1 \end{aligned} \tag{2.8}$$

Here $J_i(\tau, \mathbf{p})$ are assumed positive (Heinrich and Schuster, 1996; Kholodenko et al., 1997). Using definitions (Equation 2.5), the first-order sensitivity functions can be obtained for any model's parameter. Ranking amplitudes of $U_k(\tau, \mathbf{p})$ or averaged flux control coefficients (AFCC), important processes can be identified as in Figure 2.5. These can be used for delineating those modules for which additional genomic and chemical detail would be required.

Stability analysis shows the model's potential (via a Hopf bifurcation) for modulated quasi-periodic oscillations with large secondary period $T_2 = \frac{2\pi}{\omega_2}$, $T_2 \sim 28hr$, where $\omega_2 = Im\mu$, and μ is a complex multiplier with the largest imaginary part inside the unit circle on the complex plane depicted in Figure 2.6. We find that while metabolic processes contribute to the fastest part of the 'cell clock' with division period T of 45 min, a slower part of the clock has to be transmitted between cell generations.

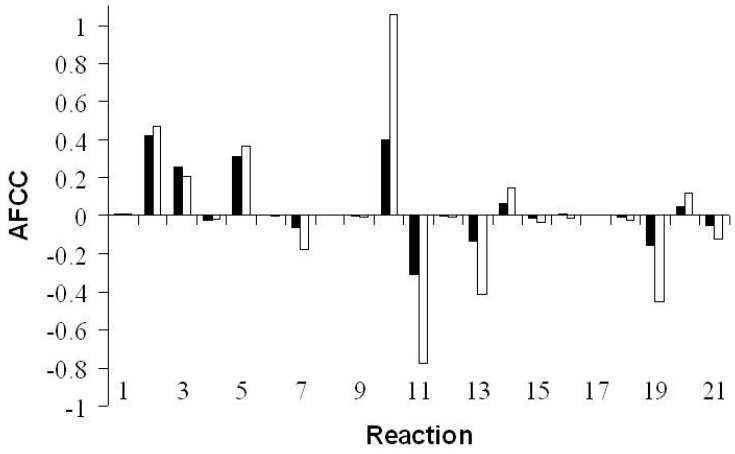


Figure 2.5: Sensitivity of the *E. coli* model to changes in parameters. Black and white bars correspond to AFCCs of the specific growth and lipids synthesis rates, respectively. The processes labeled by integer numbers are depicted in Figure 2.1.

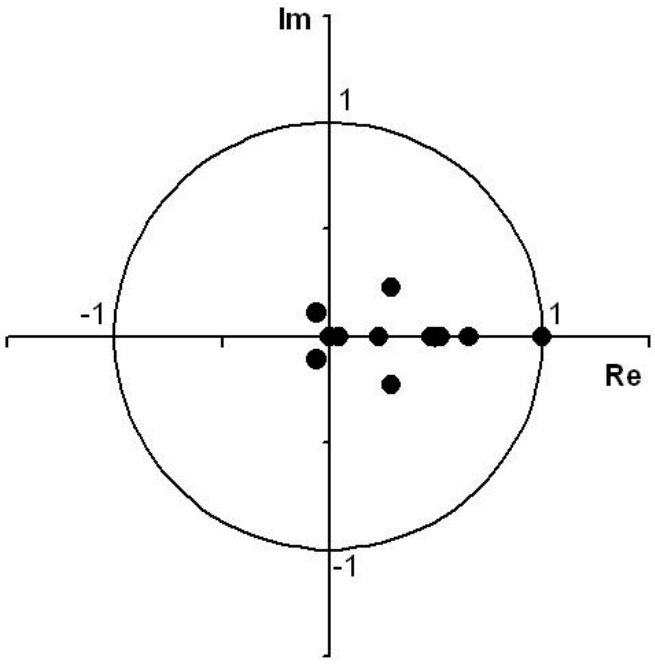


Figure 2.6: Multipliers of the limit cycle. Only a few multipliers have large magnitudes, while the others' magnitudes are very small and are clustered close to the origin.

2.5 Conclusions

We are currently developing general MCA and BVP approaches to study the robustness of hybrid whole-cell models when parameters are allowed to vary. This includes relating variations in growth conditions to changes in the number of replicating chromosomes, sensitivity analysis of the DNA replication, identification of independent measurements to fit important parameters, etc. We hope that these approaches will also help us construct large-scale genome specific modules for *E. coli* and other genomically related Gram-negative organisms (e.g. *Shewanella oneidensis* and *Zymomonas mobilis*).

REFERENCES

Doedel, E. J., Govaerts, W., and Kuznetsov, Y. (2004). Computation of periodic solution bifurcations in odes using bordered systems. *SIAM Journal of Numerical Analysis*, 41, 401–435.

Domach, M. M., Leung, S. K., Cahn, R. E., Cocks, G. G., and Shuler, M. L. (1984). Computer model for glucose-limited growth of a single cell of *Escherichia coli* B/r-A. *Biotechnology and Bioengineering*, 26(9), 1140. doi:10.1002/bit.260260925.

Heinrich, R. and Schuster, S. (1996). *The Regulation of Cellular Systems*. Chapman & Hall, Boston.

Kholodenko, B. N., Demin, O. V., and Westerhoff, H. V. (1997). Control analysis of periodic phenomena in biological systems. *Journal of Physical Chemistry B*, 101, 2070–2081.

Nikolaev, E. V., Atlas, J. C., and Shuler, M. L. (2007). Sensitivity and control analysis of periodically forced reaction networks using the Green's function method. *Journal of Theoretical Biology*, 247(3), 442–461. doi:10.1016/j.jtbi.2007.02.013.

Overbeek, R., Bartels, D., Vonstein, V., and Meyer, F. (2007). Annotation of bacterial and archaeal genomes: improving accuracy and consistency. *Chemical Reviews*, 107(8), 3431–3447. doi:10.1021/cr068308h.

Overbeek, R., Begley, T., Butler, R. M., Choudhuri, J. V., Chuang, H. Y., et al. (2005). The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Research*, 33(17), 5691–5702.

Palsson, B. O. (2004). Two-dimensional annotation of genomes. *Nature Biotechnology*, 22(10), 1218–1219.

Phipps, E. T. (2003). *Taylor Series Intergration of Differential-Algebraic Equations: Automatic Differentiation as a Toold for Simulating Rigid Body Mechanical Systems*. Ph.D. thesis, Cornell Univerity.

Shuler, M. L. (2005). Computer models of bacterial cells to integrate genomic detail with cell physiology. *Proceedings of the KBM International Symposium on Microorganisms and Human Well-Being, June 30-July 2005, Seoul Korea*.

Tomović, R. and Vukobratovitć, M. (1972). *General Sensitivity Theory*. Elsevier Publishing Co., Inc.

CHAPTER 3

INCORPORATING GENOME-WIDE DNA SEQUENCE INFORMATION INTO A DYNAMIC WHOLE-CELL MODEL OF *ESCHERICHIA COLI*: APPLICATION TO DNA REPLICATION

3.1 Abstract

The advent of thousands of annotated genomes, detailed metabolic reconstructions, and databases within the flourishing field of systems biology necessitates the development of functionally complete computer models of whole cells and cellular systems. Such models would realistically describe fundamental properties of living systems such as growth, division, and chromosome replication. This will inevitably bridge bioinformatic technologies with ongoing mathematical modeling efforts and would allow for *in silico* prediction of important dynamic physiological events. To demonstrate a potential for the anticipated merger of bioinformatic genome-wide data with a whole-cell computer model, we present here an updated version of a dynamic model of *Escherichia coli*, including a module that correctly describes the initiation and control of DNA replication by nucleoprotein DnaA-ATP molecules. Specifically, we discuss a rigorous mathematical approach used to explicitly include the genome-wide distribution of DnaA binding sites

[1]Atlas, J.C., Nikolaev, E.V., and Shuler, M.L., September 2008, "Incorporating Genome-Wide DNA Sequence Information into a Dynamic Whole-Cell Model of *Escherichia coli*: Application to DNA Replication", *IET Systems Biology*, vol. 2, no. 5, pp. 369-382, ©The Institution of Engineering and Technology 2008.

on the replicating chromosome into a computer model of a bacterial cell. We also provide a new simple deterministic approximation of the complex stochastic process of DNA replication initiation. We show for the first time that reasonable assumptions about the mechanism of DNA replication initiation can be implemented in a deterministic whole-cell model to make predictions about the timing of chromosome replication. Furthermore, we propose that a large increase in the concentration of DnaA binding boxes will result in a decreased steady-state growth rate in *E. coli*.

3.2 Introduction

3.2.1 Bacterial Cell Models

The advent of thousands of annotated genomes (Overbeek et al., 2007) and detailed metabolic reconstructions and databases (McNeil et al., 2007) in the emerging field of systems biology accentuates the need for systems level models of bacterial cells that explicitly link genomic data to fundamental properties of living systems such as growth, division, and robust control of DNA replication (Shuler, 2005). This will inevitably bridge bioinformatic technologies and data with ongoing mathematical modeling efforts and allow for *in silico* reproduction and prediction of dynamic physiological events (Palsson, 2006; Shuler, 2005; Overbeek et al., 2007). To exemplify the combination of bioinformatic genome-wide data with a whole-cell computer model, we present here an updated version of a dynamic model of *Escherichia coli*, including a module that correctly describes the initiation and control of

DNA replication by nucleoprotein DnaA-ATP molecules. Specifically, a novel rigorous mathematical approach to explicitly include genome-wide positions of the DnaA binding sites along the replicating chromosome into a computer model of a bacterial cell will be discussed. We assert that these approaches can be extended to all Gram-negative bacteria with minimal changes, including bacteria such as *Shewanella oneidensis* and *Zymomonas mobilis* which have immediate practical importance.

Many investigators have used modeling to make significant contributions to our understanding of bacterial metabolism. Some studies take advantage of detailed genomic information such as in (Keseler et al., 2005), while other models are based primarily on flux balance analysis, mathematical techniques for optimization (Palsson, 2006; Nikolaev et al., 2005), and metabolic control analysis (MCA) (Kholodenko and Westerhoff, 2004). These modeling techniques are, however, intrinsically static, and they have limited ability to predict aspects of cell regulation and dynamical response. Others have proposed methods to directly incorporate kinetic information into models of central metabolism (Chassagnole et al., 2002) or combine submodels of metabolic processes into larger cell models (Snoep et al., 2006). Some have attempted to model whole cells, for example the E-cell or Silicon Cell projects (Tomita, 2001; Tomita et al., 1997; Morgan et al., 2004). These studies, and many others, make important contributions to our perception of systems biology. However, those models often neglect important, non-metabolic aspects of cell growth (e.g. control of chromosome replication or gene duplication) because there is no formalism to handle such "events" in the context of a cell model.

The Shuler group first described a mathematical model of a single *E. coli* cell in 1979 (Shuler and Dick, 1979). At that time, it was the only model of an individual cell that did not dictate timing of cell division (e.g. growth rate) and cell size; instead, those aspects were outputs of the simulation. This "coarse-grained" model contains all of the functional elements necessary for the cell to grow, divide, and respond to a wide variety of environmental perturbations. All metabolic chemical species are included, but they are lumped into pseudochemical groups. The model is unique in its natural coupling of metabolism, transport, and cellular events, and it responds explicitly to changes in concentrations of nutrients in the environment (Domach et al., 2000). This base model has been embellished with additional biological details to allow prediction of a wide-range of responses to environmental and genetic manipulations (Shuler, 1999). The initial model included only 18 pseudochemical species that represented large groups of related metabolites. Figure 1 lists the model components and graphically depicts the relationships between them.

The mathematical description of the cellular functions in the model is based on time-variant mass balances for each component. Each mass balance takes into account the component's synthesis, utilization, and degradation, as a function of availability of precursors, energy, and relevant enzymes. Stoichiometric coefficients for relating components through mass balances were derived primarily from published research or experiments. It is important to note that the model was *not* developed by using adjustable parameters to fit model predictions to experimental results, nor did the stoichiometric mass balances assume a steady-state (i.e. the amount of each component was allowed to vary with time). Despite the simplifications made in describing the cell, the

model accurately predicts changes in cell composition, size, and shape as a function of changes in external glucose and ammonium concentration (Domach et al., 2000; Lee et al., 1984; Shuler and Domach, 1983).

The original model explicitly describes discrete events that are typically ignored in other models (Nikolaev et al., 2006). For example, changes in gene dosage (the number of copies of a gene in a cell at a given time) depend on the replication fork position, and the completeness of cross-wall formation depends on the cell size and amount of cell membrane components synthesized. Other biochemical details have been added in subsequent studies; for example, amino acids are differentiated into five families (Shu and Shuler, 1991) and the synthesis of ribosomes has been incorporated in greater detail (Laffend and Shuler, 1994a). These expansions allow the study of amino acid supplementation (Shu and Shuler, 1991) and of competition between recombinant mRNA and ribosomal mRNA in the context of high translational activity (Laffend and Shuler, 1994a). The model has also been applied to improve the use of plasmids for recombinant protein production (Laffend and Shuler, 1994a,b; Kim et al., 1987; Kim and Shuler, 1990b,a).

More recently, we have extended the classical steady-state MCA to the case of periodic processes (Nikolaev et al., 2007) to link the replicon's periodic control coefficients to the sensitivities of metabolic processes in the entire cell (Nikolaev et al., 2006). Bailey reviewed the importance of the *E. coli* model to the whole field of mathematical modeling in biochemical engineering (Bailey, 1998). The current study improves the existing model by adding mechanistic and genomic detail to the DNA replication module. This allows us to make predictions about the effect of DnaA binding site concentration on cell growth.

3.2.2 DNA Replication in Gram-Negative Bacteria

To explain the structure of the new model, we describe here some important aspects of the DNA replication process in *E. coli*. The time of initiation of DNA replication and the rate of movement of the DNA replication fork along the chromosome alter cell physiology. Figure 3.1 summarizes the major processes of DNA replication initiation control. The nucleoprotein DnaA has been shown to act as an initiator of chromosome replication. Initiation of the DNA replication process requires the binding of about 25-30 active DnaA molecules to the DNA origin, $oriC$ (Donachie and Blakely, 2003). When this happens, $oriC$ migrates and associates with the SeqA complex (Figure 3.1(a)). The $oriC$ and *dnaA* are sequestered in the SeqA-DnaA protein complex for about one-third of the cell-cycle (Figure 3.1(b)). The occupation of the *dnaA* promoter by the SeqA protein causes repression of *dnaA* expression (Torheim and Skarstad, 1999). SeqA spreads over $oriC$ by cooperative binding (Skarstad et al., 2000), releasing DnaA from $oriC$. Also, acidic phospholipids inhibit DnaA binding to $oriC$ (Figure 3.1(b)).

DnaA exists primarily in two forms in the cell: an active form bound to ATP (DnaA-ATP), and an inactive form bound to ADP (DnaA-ADP). A myriad of competing processes dynamically control the relative concentrations of DnaA-ATP and DnaA-ADP. DnaA protein is continuously produced from the *dnaA* gene and almost instantaneously assumes the DnaA-ATP nucleotide form due to the abundance of ATP molecules in the cytoplasm. Due to weak intrinsic hydrolysis, DnaA-ATP is slowly converted into DnaA-ADP. The electrostatic-hydrophobic interaction between basic DnaA molecules and acidic phospholipids facilitates the insertion of DnaA into the lipid membrane leading

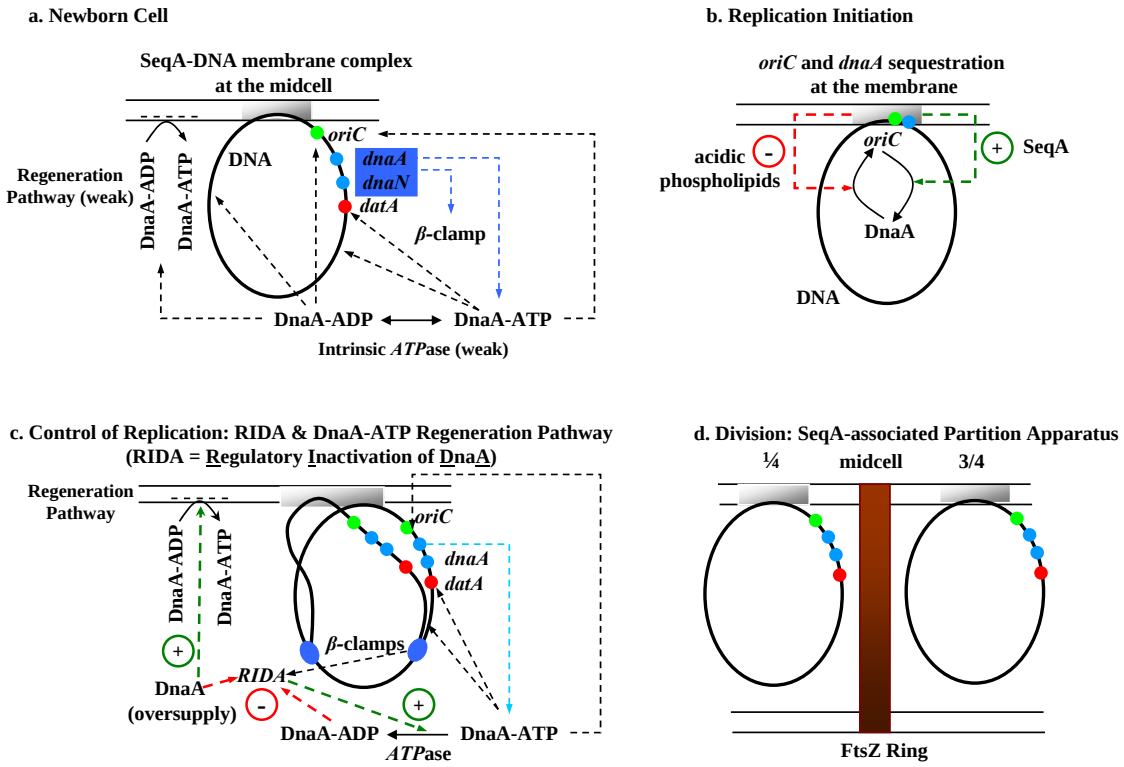


Figure 3.1: DNA replication in a Gram-negative bacterial cell. Oval rings are chromosomes. Parallel solid lines represent the cell membrane. **(a)** Newborn Cell: Control of active DnaA in a newborn cell. The *datA* box is a region that tightly binds and titrates many DnaA molecules, yet is not involved directly in initiation (Katayama et al., 2001). *dnaN* codes for the β clamp subunit of the replisome.**(b)** Replication Initiation: At the moment of replication initiation, the *oriC* is sequestered to the cell membrane for about one-third of the cell-cycle. **(c)** Control of replication: RIDA and DnaA-ATP regeneration pathway (RIDA = Regulatory Inactivation of DnaA). **(d)** Division: SeqA-associated partition apparatus model (Hiraga et al., 1998).

to its conformational change back to the DnaA-ATP form (Castuma et al., 1993; Crooke et al., 1992; Garner et al., 1998; Kitchen et al., 1999; Sekimizu and Kornberg, 1988; Yung and Kornberg, 1988). Like many peripheral proteins, DnaA is in a dynamic equilibrium between membrane-bound and soluble forms (Figure 3.1(a)).

The nucleotide forms of DnaA protein (DnaA-ATP and DnaA-ADP) are carefully controlled during the cell division cycle (Bremer and Churchward, 1991; Kurokawa et al., 1999; Messer, 2002; Speck et al., 1999). The moving β-clamp-associated Regulatory Inactivation of DnaA (RIDA) factor positively accelerates hydrolysis of ATP to ADP-bound forms to repress extra initiations (Katayama et al., 1998) (Figure 3.1(c)). The content of the ATP-bound form of DnaA protein is maintained at a low level (but not less than 100 molecules per cell) and only around the time of initiation is increased by 80% (Donachie and Blakely, 2003). Accumulation of DnaA-ATP requires efficient regeneration of DnaA-ADP to DnaA-ATP and temporal inhibition of RIDA. DnaA-ATP titration to multiple nonspecific binding sites also reduces the accumulation of free DnaA-ATP in the cell (Schaefer and Messer, 1991) (Figure 3.1(c)). The SeqA-DNA complex might act as the centromere for the chromosome, and at the time of initiation it too duplicates. One copy is subsequently passed to each daughter cell. Coincident with termination of a round of chromosome replication, these two SeqA complexes migrate in opposite directions from midcell towards the $\frac{1}{4}$ and $\frac{3}{4}$ positions. Therefore, prior to septum formation, the cell has two SeqA foci at the cell quarter sites (Hiraga et al., 1998) (Figure 3.1(d)).

A DnaA box is a DNA sequence that binds the DnaA protein, and

DnaA boxes of varying strengths are known to exist (Schaefer and Messer, 1991; Schaper and Messer, 1995). Titration of both nucleoprotein forms of DnaA protein by DnaA-binding boxes along the replicating chromosome helps control DNA replication initiation (Hansen et al., 1991b) (Figure 3.1(c)). To understand the effect of DnaA boxes on cell behavior, a more complete mechanistic description of the dynamic changes in the number of the boxes along the replicating chromosome is required. The sequence positions of the corresponding DnaA-binding sites along the chromosome are available from the organism's complete genomic sequence. Given these positions, DnaA binding sites can be directly incorporated into the model. Because the number of DnaA-ATP molecules bound to $oriC$ is relatively small, the robustness of replication with respect to stochastic fluctuation of DnaA monomers has been investigated (Browning et al., 2004). It was established that the process is robust to fluctuations, and that it can therefore be modeled using a deterministic method rather than a computationally expensive stochastic approach.

The goal of this chapter is to demonstrate the potential for explicitly merging genome-wide bioinformatic data with whole-cell modeling efforts. Specifically, we aim to directly include DNA sequence information in the chromosome replication module of the *E. coli* model described in Section 3.2.1. Previous studies have made significant advances in the modeling of DNA replication in *E. coli* (Mahaffy and Zyskind, 1989; Hansen et al., 1991b; Browning et al., 2004). In (Mahaffy and Zyskind, 1989), five states of DnaA protein were modeled, including active and inactive forms. This single-cell model included a stochastic description of the binding of DnaA to $oriC$, while the other biochemical reactions were described deterministically. Later, (Hansen et al., 1991b) introduced the concept of initiation-titration, where the free DnaA

concentration is modulated by the presence of DnaA boxes on the chromosome, into a computer model that did not include cell-division. While both of these studies did acknowledge that DnaA binds to the chromosome, and that this sequestration affects DNA initiation, neither study used sequence information when modeling the distribution of DnaA boxes on the chromosome (nor was such information available, at the time). Furthermore, these models did not address the presence of DnaA binding sites on the chromosome with varying affinity. Here, we draw on these studies to create, for the first time, a whole-cell deterministic model of DnaA-ATP controlled DNA replication in *E. coli* which takes advantage of more recent experimental discoveries. This model uses specific bioinformatic sequence information as a basis for modeling the distribution of DnaA boxes of varying affinity on the chromosome.

3.3 Methods and Model Description

3.3.1 Modeling DNA Replication Timing

The original *E. coli* model proposed that initiation of DNA replication was controlled by a hypothetical repressor protein encoded by the *dnaA* gene (Domach et al., 2000). It is now known that *dnaA* actually codes for an *initiator* protein, DnaA, that promotes DNA replication initiation (Speck et al., 1999). Many experimental observations (Donachie and Blakely, 2003; Hansen et al., 1991a; Messer, 2002; Speck et al., 1999) and computational modeling studies (Bremer and Churchward, 1991; Hansen et al., 1991b) have revealed the importance of DnaA binding boxes for determining the timing of DNA

replication initiation. Only the active nucleotide state of DnaA (i.e. DnaA-ATP) can initiate replication. The concentration of free DnaA-ATP, and therefore the timing of DNA replication initiation, is regulated by four independent mechanisms (Camara et al., 2005):

- **Initiator Titration** - The titration of newly synthesized DnaA molecules by DnaA binding boxes throughout the cell cycle (Hansen et al., 1991b).
- **Regulatory Inactivation of DnaA (RIDA)** - RIDA promotes the hydrolysis of ATP bound to DnaA, thereby deactivating it. RIDA is stimulated by DNA synthesis, resulting in a negative feedback effect which helps prevent initiation from occurring too frequently (Katayama et al., 1998).
- **Membrane Sequestration** - After initiation, the origin of replication (i.e., $oriC$) is sequestered to the SeqA protein in the cell membrane, which forces the release of DnaA molecules and prevents re-initiation for one-third of the cell cycle (Messer, 2002).
- **Semi-Methylation** - After an origin is initiated, it is unable to undergo another immediate initiation, possibly due to membrane sequestration of the incompletely methylated chromosome (Skarstad et al., 2000).

Taken together, these mechanisms prevent "false-start" initiations. Some of the essential mechanisms have been implemented in the new DNA replication module, including DnaA titration and activation. Figure 3.2 summarizes these interacting regulation processes (Camara et al., 2005).

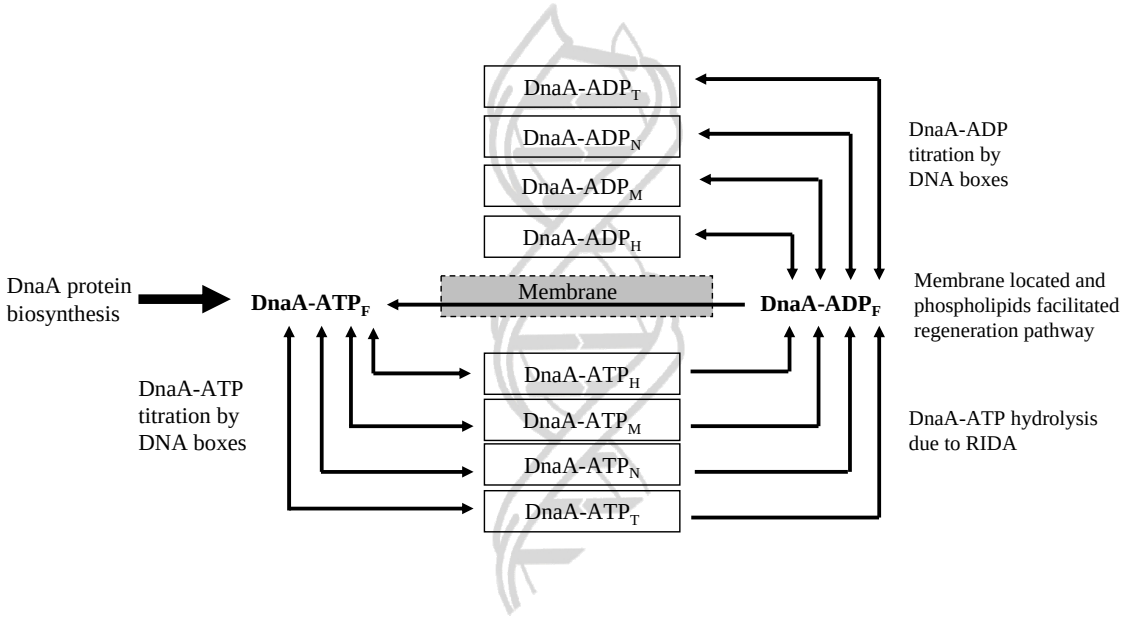


Figure 3.2: DnaA-ATP activation/inactivation and the regulation of DNA replication initiation pathways. DnaA nucleoproteins with 'F' subscripts are free in the cytoplasm. DnaA binding boxes are either High affinity (H), Medium affinity (M), or Low affinity/Nonspecific (L). T denotes the Trigger R5 DnaA box.

3.3.2 Dynamical Changes in the Number of DnaA-Binding Boxes Along the Replicating Chromosome

The following four important types of DnaA boxes and their binding affinities have been identified (Donachie and Blakely, 2003; Schaefer and Messer, 1991; Schaper and Messer, 1995):

- (H) Nine high affinity boxes.
- (M) Ninety-four medium affinity boxes.
- (L) Low affinity (nonspecific) boxes uniformly distributed along the chromosome.

- (T) Trigger box R5, which is directly involved in the initiation of the DNA replication.

In reality, the boxes display a spectrum of binding affinities which we neglect here for simplicity. The DnaA titration and DnaA-ATP initiation reaction pathways model are schematically depicted in Figure 3.2.

Note that only one molecule of nucleoprotein DnaA at a time can bind a box, while about 25-30 nucleoprotein DnaA molecules can form a complex at the chromosomal origin, $oriC$. To obtain a genome-wide distribution of the spatial positions of the DnaA-binding boxes along the bacterial chromosome, we have searched the complete *E. coli* K-12 genome in windows of 9bp corresponding to the consensus sequence TT(A/T)TNCACA (Schaper and Messer, 1995). The search algorithm simply steps through the genome one window at a time locating occurrences of the DnaA box sequence (Browning et al., 2004). The search provided us with the chromosomal positions of H and M specific boxes leading to the construction of the cumulative number distributions (*CND*) of the H and M boxes (Figure 3.3). These *CND*s are obtained by starting with the number of boxes near the DNA terminus and adding each additional box on the chromosome as one follows along the DNA up to the *oriC* position. In statistics, similar distributions are referred to as *cumulative frequency distributions*. Because the number of the nonspecific boxes (i.e. L-boxes) should be proportional to the total chromosomal length, we approximate its *CND* using a uniform distribution described by the scalar factor a_L, where a_L is the total number of all *E. coli* DNA base pairs, N_{bp} = 4639221 (Blattner et al., 1997), divided by 9 (the length of the consensus sequence) (3.1).

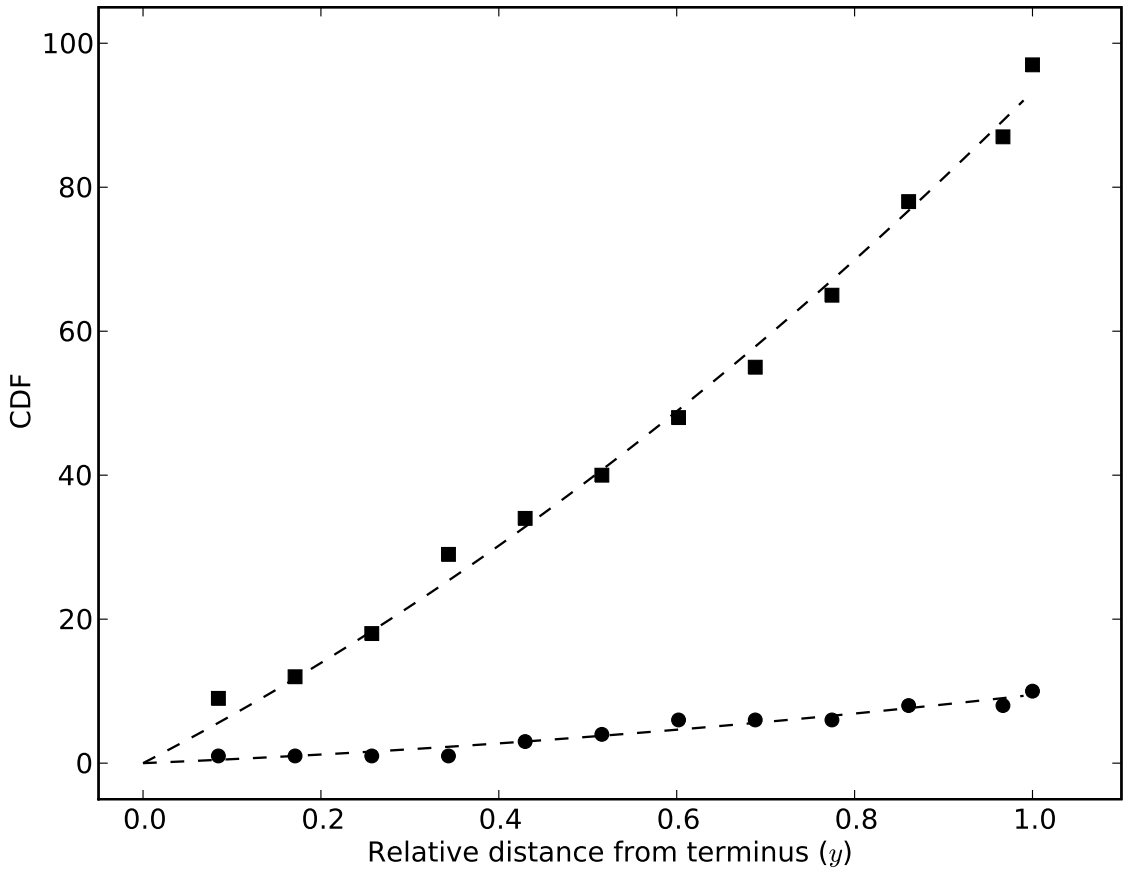


Figure 3.3: Cumulative number distribution functions (*CND*s) for the high (i.e. H) and medium (i.e. M) affinity DnaA-binding boxes along the *E. coli* K-12 chromosome. The circles correspond to the high affinity H-boxes (i.e. *CND* is $F_1(y)$) and the squares correspond to the medium affinity M-boxes (i.e. *CND* is $F_2(y)$). Coordinate y is the fractional distance along the chromosome counted from its terminus. The dashed curves correspond to the best fit to equation (3.2).

$$a_L = \frac{N_{bp}}{9} = 515496 \tag{3.1}$$

The *CND*s for the H-, and M- can be fitted using the quadratic form in equation (3.2).

$$F_k(y) = a_k \cdot y + b_k \cdot y^2, (k = H, M, L) \tag{3.2}$$

Here the distribution parameters have been fit in Figure 3.3, with a_H = 63.694 and b_H = 29.596 for H-boxes, a_M = 5.1201 and b_M = 4.368 for M-boxes, and a_L = 515496 and b_L = 0 for L-boxes. Coordinate y in (3.2) is the fractional distance along half of the chromosome counted from its terminus, such that $y = 0$ corresponds to the terminus and $y = 1$ corresponds to $oriC$. Coordinate y is counted from the terminus rather than from $oriC$ because the chromosomal origin can replicate multiple times for a single terminus. After initiation, two replicating forks progress along the chromosome with the same rate in opposite directions from $y = 1$ to $y = 0$. Because the forks move at the same rate, we can consider only half of the circular chromosome (Domach et al., 2000).

The expression (3.2) and the corresponding *CND*s have been obtained by recalculating the positions of all DnaA boxes in terms of the y-coordinate, and then fitting *CND*s of the form of (3.2) to the distributions. Using *CND*s is possible because we are only interested in the timing of appearance of the newly synthesized DnaA boxes and not in their absolute spatial positions along the entire chromosome. The nonlinear cumulative distributions (3.2) can significantly contribute to robust DNA replication initiation control (Hansen

et al., 1991b).

It is important to note that a_H, b_H, a_M, and b_M are parameters that describe precisely the distribution of DnaA boxes in the *E. coli* genome, while we are postulating that parameter a_L alone can describe the theoretical upper limit on the distribution of non-functional binding sites on the chromosome. We consider the effect of varying the H-, M-, and L- box distributions in Section 3.4.1, where the H- and M- box concentrations can be increased or decreased, but the L-box concentration can only be decreased.

A growing *E. coli* cell can have up to 14 replication forks moving along the chromosome simultaneously (Figure 3.4). Although a pair of forks is always assembled on the original chromosome (Figure 3.4(a)), there can be two more pairs of moving forks synchronously emanating from the two new $oriC$ (Figure 3.4(b)). Similarly, there can be four new pairs of moving forks synchronously initiated before the previously initiated forks reach the terminus (Figure 3.4(c)). Given the complexity of the DNA replication dynamical process, it is important to rigorously describe the dynamical changes in the corresponding DnaA-binding boxes along the replicating DNA strand. Pairs of moving forks, simultaneously emanating from the same $oriC$, can be described in terms of a single coordinate position y along the chromosome. We denote such representative forks as $Fork_1$, $Fork_2$ and $Fork_3$, which have coordinates y_1, y_2, and y_3, respectively, such that $0 \leq y_1 < y_2 < y_3 \leq 1$ and $y = 0$ corresponds to the terminus. Then, using the *CND*s given by the expressions in (3.2), the dynamical changes in the numbers of H-, M-, and L-binding boxes can be calculated by equations (3.3), (3.4), and (3.5) (see Appendix H).

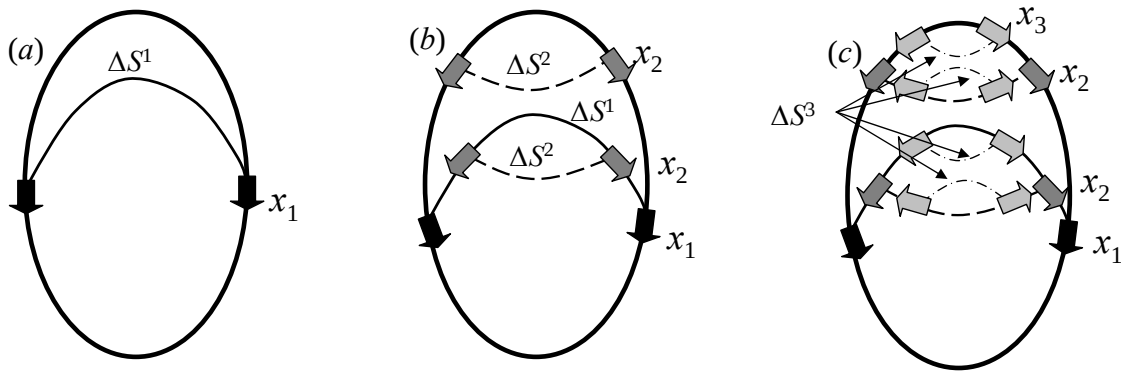


Figure 3.4: Replication fork counting. Depending on the external environment, a growing *E. coli* cell can have (a) 2, (b) 6 and (c) 14 replication forks moving along the replicating chromosome. ΔS is the fraction of the DnaA-binding boxes formed on the newly synthesized lagging strands, (a) $\Delta S = \Delta S^1$, (b) $\Delta S = \Delta S^1 + 2\Delta S^2$, and (c) $\Delta S = \Delta S^1 + 2\Delta S^2 + 4\Delta S^3$.

$$S_k = N_{chrom} \cdot (a_k \cdot A(y_1, y_2, y_3) + b_k \cdot B(y_1, y_2, y_3)), (k = H, M, L), \tag{3.3}$$

$$A(y_1, y_2, y_3) = y_1 + 2(y_2 - y_1) + 4(y_3 - y_2) + 8(1 - y_3), \tag{3.4}$$

$$B(y_1, y_2, y_3) = y_1^2 + 2(y_2^2 - y_1^2) + 4(y_3^2 - y_2^2) + 8(1 - y_3^2). \tag{3.5}$$

Here S_H is the number of high affinity H-boxes, S_M is the number of medium affinity M-boxes, and S_L is the number of nonspecific low affinity L-boxes. N_{chrom} is the total number of synchronously replicating chromosomes, $N_{chrom} \in \{1, 2, 3\}$. Function $A(y_1, y_2, y_3)$ represents the total length of the symmetric half of the replicating chromosome, while function $B(y_1, y_2, y_3)$ is a nonlinear function used to calculate the total number of binding boxes. A and B are the same for all boxes, and similar functions could be applied for alternate binding boxes distributed through the chromosome. After the completion of DNA replication, defined by the time moment when $Fork_1$ reaches the terminus (i.e. when $y_1 = 0$), the Forks and their positions are updated using the update rules in (3.6) and (3.7).

$$Fork_2 \to Fork_1, Fork_3 \to Fork_2 \tag{3.6}$$

$$y_3 \to y_2, y_2 \to y_1, 1 \to y_3 \tag{3.7}$$

When DNA replication completes, the oldest replication fork "becomes" the new first fork, and the position variables y_i are likewise updated.

3.3.3 Ordered and Sequential Binding of DnaA-ATP Molecules to $oriC$

By footprint analysis and electron microscopy, it was shown that the buildup of approximately 25-30 monomers of the nucleoprotein complex of DnaA protein at the chromosomal origin (i.e. $oriC$), DnaA-ATP, is required to unwind $oriC$, resulting in DNA replication initiation (Donachie and Blakely, 2003). This DnaA-ATP binding process proceeds through seven distinct ordered states (Crooke et al., 1993; Margulies and Kaguni, 1996). Although this important experimental observation lacks explicit mechanistic molecular detail, we model the overall stochastic process by using a deterministic approximation that allows us to capture essential transitions between the seven stages of the formation of the active DnaA-ATP nucleoprotein complex at $oriC$, called here a *replicon* (Margulies and Kaguni, 1996). The formation of different oriC-DnaA protein complexes through seven distinct stages can be presumably explained by different affinity properties of the 9-mer binding sequences in $oriC$, called the R1-R4 binding boxes. Specifically, DnaA protein binds to box R4 with about 3-fold higher affinity than it binds to box R1 (Margulies and Kaguni, 1996). Therefore, the entire replicon complex is formed when DnaA boxes with higher affinities are first occupied with DnaA protein molecules which then sequester binding DnaA molecules to nearby boxes through cooperative effects (Margulies and Kaguni, 1996).

We assume inert binding of DnaA-ADP nucleotide form of DnaA protein to $oriC$ can be neglected (Crooke et al., 1993; Margulies and Kaguni, 1996). Because the number of DnaA-ATP molecules necessary to build the DnaA-ATP complex (i.e. replicon) at $oriC$ is small, a stochastic "birth-and-death" process

can formally be used to describe the corresponding complex transitions (Feller, 1968). There are four different kinds of events which change the configuration of the replicon at $oriC$: (i) spontaneous association of DnaA-ATP molecules with the bare $oriC$, (ii) spontaneous association of DnaA-ATP molecules with the replicon formed at $oriC$, (iii) spontaneous dissociation of DnaA-ATP molecules from the replicon, and (iv) the spontaneous transition of the replicon between different states. The master equation describing the continuous Markov chain corresponding to the stochastic process is an infinite system of ordinary differential equations. In some cases, the unique solution to the master equation can be found using generating functions (Feller, 1968). However, in a general case, solving the master equation is complex, and reasonable approximations are needed. Here we use a simple deterministic approximation approach to model the stochastic process of the formation and the transition of the replicon at $oriC$.

About 28 molecules of DnaA-ATP can in average bind to $oriC$ through the seven distinct ordered states required to start DNA replication in *E. coli* (Crooke et al., 1993; Margulies and Kaguni, 1996). Given this experimental evidence, we assume that about four DnaA-ATP molecules can in average bind to the replicon at $oriC$ at each of the seven replicon states. Let P_{oriC} be a stationary probability that the replicon moves between states R and $R+1$, $R = 0, 1, \ldots, 7$. Once $R = 7$, four more molecules of DnaA-ATP bound to the replicon will trigger the initiation of DNA replication. We denote by $\langle N^R_{DnaA} \rangle$ the averaged number of DnaA-ATP molecules bound to the replicon at state R. Then $\langle N^R_{DnaA} \rangle$ is proportional to the number of DnaA binding boxes N_B = 4, times the replicon state (i.e. R) times the probability that the replicon can be found at state R (i.e. $(P_{oriC})^R$). Therefore, the averaged number of DnaA-ATP molecules bound to

the replicon at state R can be approximately estimated using the mathematical expectation as in equation (3.8).

$$\langle N^R_{DnaA} \rangle \sim N_{oriC} \cdot N_B \cdot R \cdot (P_{oriC})^R \tag{3.8}$$

Here N_{oriC} is the number of all replicating DNA origins in the cell and N_B is the number of functional DnaA binding boxes within $oriC$, $N_B = 4$ (Crooke et al., 1993; Margulies and Kaguni, 1996). After binding of four more DnaA-ATP molecules to the replicon, as discussed above, $\langle N^8_{DnaA} \rangle \sim 28$ at $R = 8$. Here $R = 8$ does not correspond to any replicon state and, instead, corresponds to the discrete replication initiation event. Using $\langle N^8_{DnaA} \rangle \sim 28$, $N_{oriC} = 1$, $N_B = 4$, and $R = 8$ in (3.8), we can solve (3.8) for the probability P_{oriC}, yielding $P_{oriC} = 0.985$ used in the model. Letting $N_{oriC} = 1$ for simplicity, it can be seen from (3.8) that about four DnaA-ATP molecules are added to the replicon after each transition $R \to R + 1$.

Let P_t be a monotonically increasing function that can be interpreted as a "replicon state transition probability" at time t. Then the time moment $t = t'$ corresponding to the actual discrete event of the transition between the replicon states R and $R + 1$ can be determined by the event condition (3.9).

$$P_{t'} = P_{oriC} \tag{3.9}$$

Therefore, for $t'' < t < t'$, with t'' corresponding to the time of the previous event transition, no transitional event is possible (i.e. because $P_t < P_{oriC}$). Let $\overline{S}_{DnaA}$ be the number of the DnaA-ATP-bound H-boxes outside the formed replicon and let $\overline{S}_H$ be the number of free H-boxes at time t. Then the monotonically

increasing transitional probability P_t can be estimated at time t using the combinatorial formula (3.10) as discussed in Appendix H.

$$P_t \sim \frac{\Gamma(\overline{S}_{DnaA}+1)}{\Gamma(\overline{S}_{DnaA}-N_B+1)} \cdot \frac{\Gamma(\overline{S}_H-N_B+1)}{\Gamma(\overline{S}_H+1)} \tag{3.10}$$

Here $\Gamma(n)$ is the Gamma function, which for integer values of n is $\Gamma(n+1) = n!$ (W.H. et al., 1988). The probability of the formation of the replicon at bare $oriC$ can be obtained in a similar way (see Appendix H). The nucleoprotein DnaA is continually synthesized, causing a corresponding increase in the number of DnaA molecules bound to H-boxes outside the replicon (i.e. $\overline{S}_{DnaA}$), along with a decrease in the number of free H-boxes (i.e. $\overline{S}_H$). We find from (3.10) that overall this process increases the chance (i.e. P_t) that the next N_B DnaA-ATP molecules (i.e. $N_B = 4$) will bind to the replicon at $oriC$, resulting in the transition of the replicon to the next state.

3.3.4 Coupling the DNA Replication Module to the Whole-Cell Model

We couple the model of DNA replication with the previously developed whole-cell model through the following four key dynamical processes which are schematically depicted in Figure 3.2 and discussed below:

1. The rate at which replication forks move along the DNA molecule, which is proportional to the rate of M_3 synthesis in Figure 3.5.
2. The active DnaA protein is produced by constitutive protein synthesis

(proportional to the rate of synthesis of M_1 in Figure 3.5).

3. The active DnaA-ATP protein is regenerated by membrane phospholipids with acidic head groups, which can catalyze the rapid release of nucleotides from DnaA, rejuvenating the ATP form from the ADP form (dependent on the value of P_4 in Figure 3.5).

4. We link the inactivation of DnaA-ATP by conversion to DnaA-ADP due to the RIDA process to time-dependent changes in M_3 (Figure 3.5).

When writing the rate equations for the new module, we follow a general approach of the approximation of reaction rates and the selection of model's parameters previously introduced and discussed in the works (Domach et al., 2000; Browning et al., 2004). The detailed descriptions of reaction rates and kinetic parameters are encoded in the model's SBML representation, which is available upon request.

(1) We begin with the mathematical description of the moving replication forks. Let $x_k = 1 - y_k$ be defined as the position of the k^{th} fork, with respect to $oriC$. We first note that concurrent initiation of DNA replication occurs at all $oriC$ sites present in the cell once per cell cycle (Kitagawa et al., 1998). Therefore, the dynamical changes in coordinate x_k moving along the replicating chromosome can be modeled by linking the monotonic changes in $x_k(t)$ to the monotonic change in the fraction of newly synthesized DNA mass (i.e. $M_{DNA}(t)$) per replicating chromosome using rate law (3.11).

$$\frac{dx_i}{dt} = \frac{1}{N_{tot}} \cdot \frac{1}{M_{DNA}} \cdot \frac{dM_{DNA}}{dt}, (i = 1, 2, 3) \tag{3.11}$$

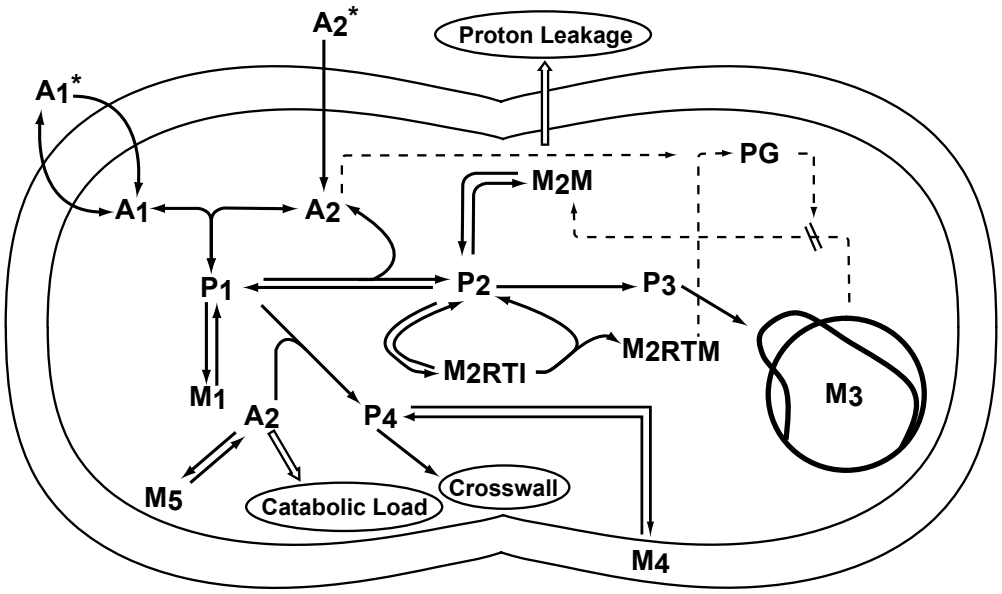


Figure 3.5: General overview of the Cornell *Escherichia coli* model. Note: Not all reactions and regulation information are depicted. Species M_3 is the replicating chromosome. For a detailed discussion of the coupling between the original *E. coli* model and the new DNA replication module, please see Section 3.3.4 and Figs. 3.1 and 3.2. Solid lines represent pseudochemical reactions. Dashed lines represent the flow of information. Other species: A_1 - ammonium ion, A_2 - glucose, P_1 - amino acids, P_2 - ribonucleotides, P_3 - deoxyribonucleotides, P_4 - membrane precursors, M_1 - protein, M_{2RTI} = immature stable RNA, M_{2RTM} - mature stable RNA, M_3 - DNA, M_4 - cell envelope, M_5 - glycogen, PG - ppGpp, E_1 - enzymes for conversion of P_2 to P_3, E_2 & E_3 - enzymes for cross-wall formation and cell envelope synthesis. * indicates species that are external to the cell. Figure adapted from (Domach et al., 2000; Nikolaev et al., 2006).

Here t is time, N_{tot} is the total number of replicating forks, M_{DNA} is the mass of the replicating chromosomes, and dM_{DNA}/dt is the rate of DNA biosynthesis as described in (Domach et al., 2000).

(2) We model the rate of DnaA biosynthesis (V_{DnaA}) by coupling it with the rate of the total protein biosynthesis (i.e $(dM_1/dt)_S$) as in (3.12):

$$V_{DnaA} = \frac{k_{DnaA} \cdot f_{GD}}{(1 + \alpha \cdot \frac{N_{ATP,M}}{N_M} + \beta \cdot \frac{N_{ADP,M}}{N_M})} \cdot (\frac{dM_1}{dt})_S \tag{3.12}$$

In (3.12), k_{DnaA} is the kinetic rate constant for DnaA synthesis (Browning et al., 2004). Parameters α and β represent noncompetitive autorepression of DnaA biosynthesis as discussed earlier. We use $\alpha = 2$ and $\beta = 0.02$ (Browning et al., 2004). Ratios $N_{ATP,M}/N_M$ and $N_{ADP,M}/N_M$ are the fractions of medium affinity DnaA boxes occupied by DnaA-ATP and DnaA-ADP, respectively. The formation of mRNA transcripts necessary for the biosynthesis of total protein (i.e. M_1) was discussed in detail in our previous work (Domach et al., 2000; Laffend and Shuler, 1994b). The time-dependent value of f_{GD} in (3.12) is defined by formulas:

$$f_{GD} = \frac{dnaA_{GD}}{Total_{genes}}, \tag{3.13}$$

$$Total_{genes} = N_{chrom} \cdot DNA_{genes} \cdot (1 + Fork_1 + 2 \cdot Fork_2 + 4 \cdot Fork_3) \tag{3.14}$$

Here $dnaA_{GD}$ is the total dnaA gene dosage calculated for all replicating chromosomes at time t, $Total_{genes}$ is the total number of all genes in all N_{chrom}

synchronously replicating chromosomes, DNA_{genes} is the number of genes in one *E.coli* chromosome, $DNA_{Genes} = 4405$.

(3) The rate expressions for the membrane-mediated regeneration of free DnaA-ATP from free DnaA-ADP is given by (3.15):

$$V_{reg} = \frac{k_{reg} \cdot P_4/V}{(K_{reg} + P_4/V)} \cdot DnaA_{ADP,F} \tag{3.15}$$

Here $DnaA_{ADP,F}$ is the number of free DnaA-ADP molecules (i.e. not bound to DNA), k_{reg} is the kinetic regeneration rate constant, K_{reg} is the saturation constant for membrane lipids, and P_4/V is the cellular concentration of envelope precursors (Domach et al., 2000).

(4) We describe the rate of RIDA mediated inactivation of DnaA-ATP molecules using the formula:

$$V_{inact} = \frac{k_{inact}}{(1 + DnaA_{ADP,F}/K_{inact})} \cdot \frac{dM_1}{dt}. \tag{3.16}$$

Here k_{inact} is the kinetic rate constant of the RIDA inactivation, K_{inact} is the noncompetitive inhibition constant for RIDA by free DnaA-ADP molecules.

3.3.5 Model Implementation and Simulation

The updated model is available in the Systems Biology Markup Language (SBML). The SBML version of the model contains 33 species, 42 reactions, and over 30 discrete events. Model simulations reported here were performed using SloppyCell (Gutenkunst et al., 2007), a software environment for simulation and analysis of biomolecular networks written in the Python programming

language. All model simulation results presented here are generated by integrating the model from an initial condition until a stable cell-division *limit* cycle is reached (Nikolaev et al., 2006). It is common to study how bacterial behavior changes at different steady-state growth rates, which is controlled by varying the external nutrient concentration. In the simulation results presented in Figure 3.6, where growth rate is varied, the actual control parameter is the external glucose concentration. Growth rate can also be used as a reporter of the effect of varying a particular parameter, as in Figure 3.7. The original model corresponds to a single cell of *E. coli* B/r growing at steady-state in a constant chemical environment (Domach et al., 2000), but it has been compared to other strains and posed as a generalized model of a *chemoheterotrophic* bacterial cell (Browning and Shuler, 2001; Nikolaev et al., 2006).

3.4 Results and Discussion

Here we compare the model predictions to experimental data from the literature and other models. The current base model is for glucose limited growth of *E. coli* B/rA growing in a constant chemical environment (i.e. a *chemostat*), and can achieve growth rates up to $1.0hr^{-1}$. Unfortunately, it is difficult to find extensive data for chemostat cultures, as most experiments are performed in batch culture. The exponential phase of batch culture is analogous to steady-state continuous culture, as both modes correspond to *balanced growth*, where all the population averages of the chemical species in the culture increase at the same rate. The best available data for comparison is often not run at precisely the same conditions as the model. Specifically, the growth rate is varied by changing the carbon source, and via nutrient supplementation (e.g. with yeast extract), whereas in the *E. coli*

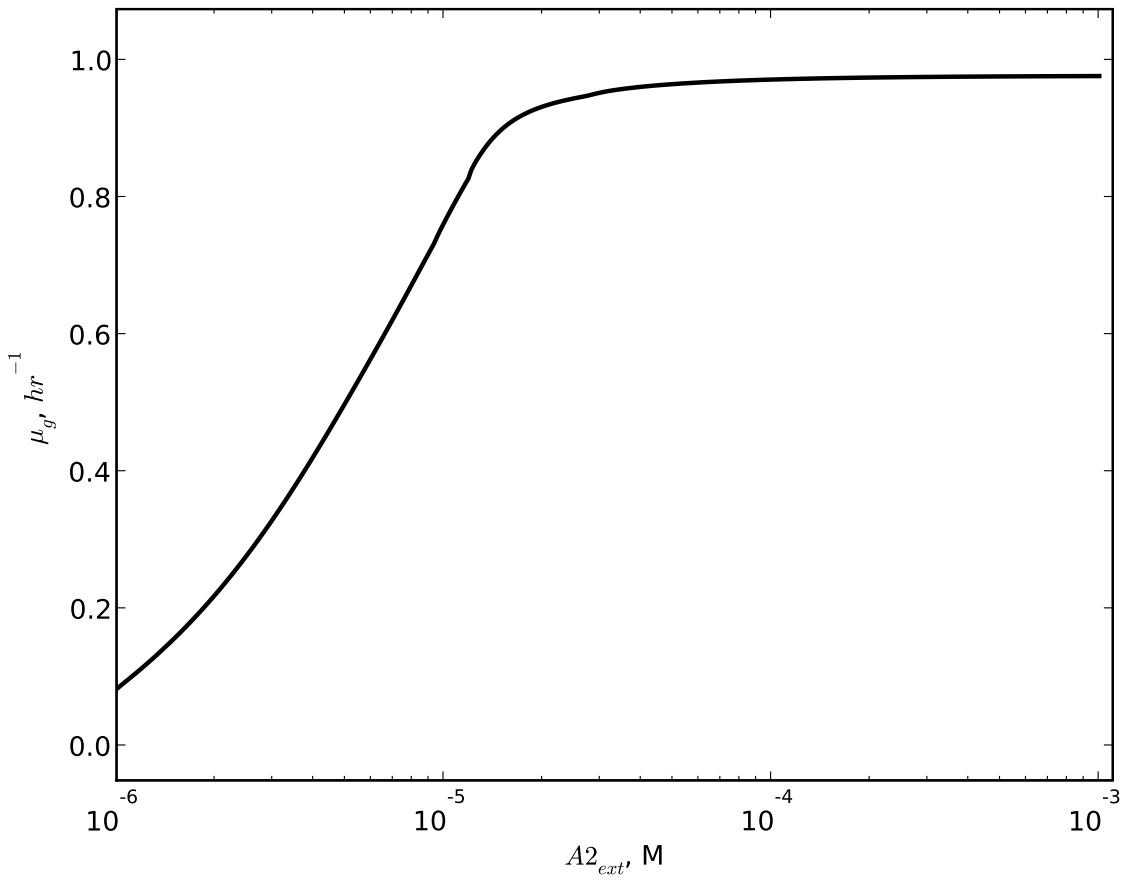


Figure 3.6: Growth rate, μ_g, vs. external glucose concentration, $A2_{ext}$. The growth rate can be directly controlled using the external glucose concentration, which is a control variable in the laboratory.

model, growth rate it is varied by changing the concentration of glucose as the sole carbon source. We make the comparison to this data in Sections 3.4.2 and 3.4.3 with the caveat that the conditions do not correspond exactly to those of the model, but some sensible conclusions can still be inferred in regard to DNA replication initiation using growth rate as a reporting variable.

3.4.1 Cell Growth Rate as a Function of DnaA Binding Box Concentration

The matter of primary importance for a whole-cell bacterial model is that it predicts a stable cell-division cycle for a variety of input conditions. Figure 3.6 shows the growth rate as a function of input external glucose concentration. Each point on the curve in this figure represents a separate model simulation, where the external glucose concentration is set, and then a steady-state cell division cycle is achieved, such that the average cell properties (e.g. growth rate) can be calculated. By calculating the growth rate for a range of values of a particular parameter, we can quickly evaluate how that parameter affects the cell.

To evaluate the importance of the DnaA box concentration along the chromosome, we performed simulations where the DnaA Box distributions (3.2) are scaled over a logarithmic range. This scales the total number of boxes available to bind DnaA, while maintaining the same quadratic shape to the distribution. The simulation results are shown in Figure 3.7, where we report results in terms of the total number of boxes in a scaled distribution. Reducing the number of medium (M) and low (L) affinity boxes down to 1 has very little

effect on the cell's growth rate. Reducing the number of high affinity (H) boxes, however, causes a disruption in the cell's ability to achieve a stable growth rate. If the concentration of boxes is instead increased, we see that for the H-, and M-boxes the growth rate plateaus and then reaches a slightly higher maximum, before dropping as the box concentration is increased. Over the range of L-box concentrations considered, there is no observable change in the growth rate. Recall that the total number of H-, M-, and L-boxes found in *E. coli* are 9, 94, and 515496, respectively (see Section 3.3.2). We observe that nature has selected an H-box concentration just above the minimum for which a stable growth cycle is achievable. The cell sees no benefit for moderate increases in the H-box concentration above 9 total boxes. A very large increase in the box concentration for H- or M-boxes can actually have a negative impact on cell growth. This is because the extra binding boxes actually titrate DnaA away from the cytoplasm so frequently that it is not present in sufficient quantities to initiate replication. While it would be challenging to introduce a large number of DnaA boxes into the *E. coli* chromosome, it should be possible to introduce a high-copy number plasmid into the cell with many copies of the DnaA binding boxes. This method could be used to experimentally confirm the effect of additional titration on DnaA mediated replication initiation.

3.4.2 DNA Replication Timing

Figure 3.8 shows the model prediction of the length of the C period (the time required for the DNA replication fork to proceed from the $oriC$ to the terminus) for our *E. coli* model with the new deterministic DNA replication module compared to a variety of *E. coli* data compiled by (Helmstetter, 1996). Predicting

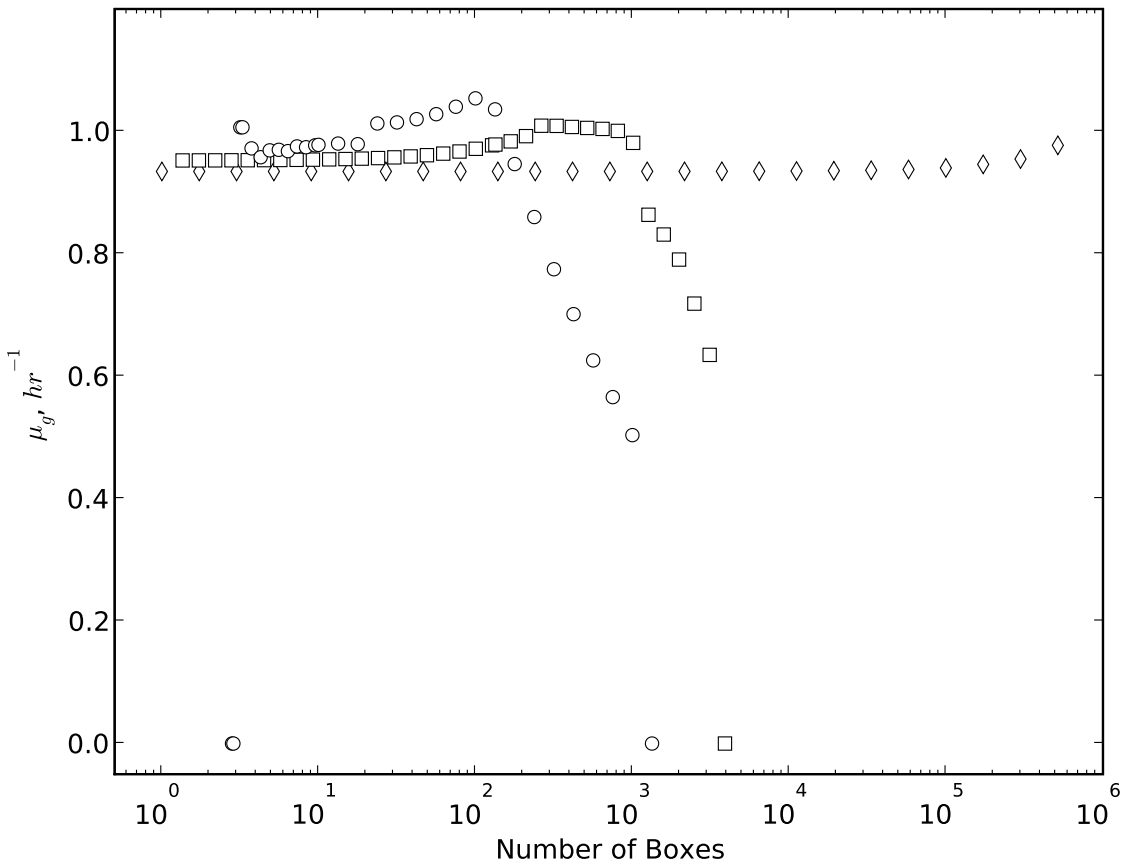


Figure 3.7: Growth rate, μ_g, vs. the total number of DnaA binding boxes for high-affinity (H-Boxes), medium-affinity (M-Boxes), and low-affinity (L-Boxes) DnaA binding boxes. ○: H-Boxes, □: M-Boxes, ♢: L-Boxes.

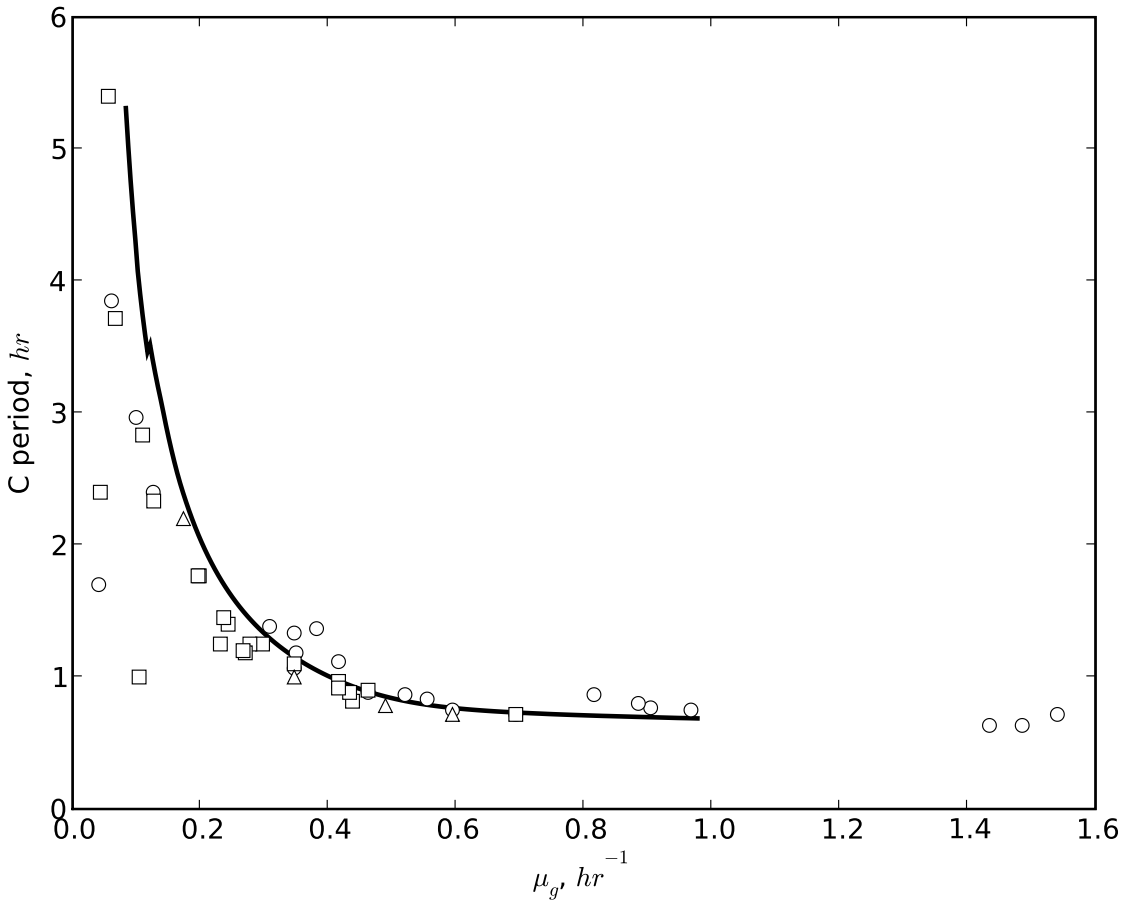


Figure 3.8: C period vs. growth rate. Data comes from a variety of sources compiled in (Helmstetter, 1996). Circles - data compiled for *E. coli* Br/A strains. Squares - data compiled for *E. coli* Br/K strains. Triangles - data compiled for *E. coli* Br/F strains.

a whole-cell model capture this behavior.

There are conflicting reports about the relation between cell size and the timing of initiation. The term 'initiation mass' was introduced as a way to parameterize the state of the cell at initiation (Donachie, 1968; Mahaffy and Zyskind, 1989). Early experiments showed that the cell will initiate DNA replication at a nearly constant initiation mass per number of origins, except at low growth rates (Donachie, 1968; Mahaffy and Zyskind, 1989). This is

due to a still unknown mechanism (Herrick et al., 1996). Other groups have reported that the initiation mass does vary continuously with growth rate at growth rates below 1.0 hr^{-1} (Herrick et al., 1996; Churchward et al., 1981). However, there has been evidence for the opposite trend, with a monotonically *decreasing* initiation mass as the growth rate is increased (Wold et al., 1994). The discrepancy in experimental data could be due to one of two factors. First, the experimental growth rate is varied by changing nutrient supplements (e.g. yeast extract) rather than by varying the glucose concentration alone, as in our model. This is a common way to vary the growth rate in culture, but it means that multiple control factors are being varied simultaneously. Secondly, it is possible that a correlation between the initiation mass and replication phase entry was established, rather than a causative relationship (Boye and Nordstrm, 2003). To our knowledge the initiation mass in *E. coli* has not been measured in continuous culture using only the glucose concentration as a control variable, which would be the most direct experimental analogue to the model presented here. Our model is based on reasonable mechanisms for DNA replication, and predicts that the initiation mass per origin increases approximately threefold as the growth rate increases steadily from $0.2hr^{-1}$ to $1.0hr^{-1}$ through glucose control. This agrees reasonably well with Figure 1 of Churchward *et al.* (Churchward et al., 1981), which predicts that in the *E. coli* B/r strain the initiation mass will nearly double as the growth rate increases from $\sim 0.3hr^{-1}$ to $\sim 1.0hr^{-1}$. This observation supports the idea that while initiation mass is a useful tool for parameterizing initiation timing, it does not directly control it.

3.4.3 DnaA Concentration

The model prediction of average DnaA content at varying growth rates was compared to the experimental data from Hansen *et al.* (Hansen et al., 1991a) and Chiaramello *et al.* (Chiaramello and Zyskind, 1989). At a low growth rate ($\sim 0.4 hr^{-1}$), the model predicts an average total DnaA concentration of 600 monomers/cell, which overestimates the corresponding experimental data measurements of 330 (Hansen et al., 1991a) and 74 (Chiaramello and Zyskind, 1989) monomers/cell. However, at a higher growth rate ($\sim 1.0 hr^{-1}$), the model predicts an average DnaA concentration of 850 monomers/cell, which better matches the experimental measurements of 700 (Hansen et al., 1991a) and 803 (Chiaramello and Zyskind, 1989) monomers/cell. Overall, our model predictions better match those in Table 2 of Hansen *et al.* (Hansen et al., 1991a), where they explain that in contrast to the experiments done by Chiaramello *et al.*, their data was collected in cultures that had been in steady-state exponential growth for more than 10 generations (Hansen et al., 1991a). Similarly, our model represents cells growing in balanced growth conditions for many generations (i.e. steady-state), and we find it striking that the data measured after a longer period in exponential growth falls closer to our model predictions.

To our knowledge a study where DnaA concentrations are measured in a steady-state chemostat has not been performed; however, we find the behavior of free DnaA-ATP during the division cycle in the deterministic model is in qualitative agreement with the results of Browning *et al.* (Browning et al., 2004). Specifically, after the initiation of replication, the number of free DnaA monomers per cell rapidly increases as they are flushed off of $oriC$. This is followed by a decrease in free DnaA-ATP (i.e. an *eclipse* period) due to

increased binding to the chromosome as the DnaA boxes are replicated. Both the deterministic model presented here and the stochastic model in (Browning et al., 2004) predict similar DnaA dynamics.

3.5 Conclusions

Synthetic biology asks the experimental question of what we can manipulate in a cell, and how the organism will respond to those manipulations. To help establish a method for answering this question computationally, we show here, for the first time, that a deterministic model of DNA replication in *E. coli* can be constructed that both incorporates explicit genomic data, and is integrated into a computer model that accounts for metabolism, cell expansion, and cell division. Other deterministic models of DNA replication are not integrated into a complete cell model that responds explicitly to changes in nutrients. The model presented here demonstrates one way to explicitly link DNA sequence information to systemic physiological behavior. Such a link is essential for the progress of synthetic biology. Our model suggests, for example, that the concentration of DNA binding boxes on the chromosome is critical to determining cell growth and behavior. We propose that by introducing a high copy plasmid with DnaA binding boxes, that the growth rate of *E. coli* may decrease due to an overwhelming draw on the free DnaA protein.

Through studying models that simulate the link between the genome and physiology, we can not only test our existing hypotheses about cellular biology, but also make novel predictions that make use of the now abundant resources of bioinformatics. The predictions of this model are nearly identical

to those from our previous model using a stochastic description of DNA replication (Browning et al., 2004). Because the deterministic model is less computationally expensive it will be preferred for most applications. Key factors in the success of the deterministic model are the natural robustness of the replication mechanism and the use of the appropriate monotonic function to correctly identify the moment of replicon formation. Further, this model demonstrates the hybrid-modular nature of this modeling approach as described elsewhere (Shuler and Domach, 1983; Nikolaev et al., 2006; Shuler, 2005; Castellanos et al., 2004).

REFERENCES

Atlas, J. C., Nikolaev, E. V., Browning, S. T., and Shuler, M. L. (2008). Incorporating genome-wide DNA sequence information into a dynamic whole-cell model of *Escherichia coli*: application to DNA replication. *IET Syst Biol*, 2(5), 369–382. doi:10.1049/iet-syb:20070079.

Bailey, J. E. (1998). Mathematical modeling and analysis in biochemical engineering: past accomplishments and future opportunities. *Biotechnology Progress*, 14(1), 8–20. doi:10.1021/bp9701269.

Blattner, F. R., Plunkett, G., Bloch, C. A., Perna, N. T., Burland, V., et al. (1997). The complete genome sequence of *Escherichia coli* K-12. *Science*, 277(5331), 1453–1474.

Boye, E. and Nordstrm, K. (2003). Coupling the cell cycle to cell growth. *EMBO Reports*, 4(8), 757–760. doi:10.1038/sj.embor.embor895.

Bremer, H. and Churchward, G. (1991). Control of cyclic chromosome replication in *Escherichia coli*. *Microbiological Reviews*, 55(3), 459–475.

Browning, S. T., Castellanos, M., and Shuler, M. L. (2004). Robust control of initiation of prokaryotic chromosome replication: essential considerations for a minimal cell. *Biotechnology and Bioengineering*, 88(5), 575–584. doi:10.1002/bit.20223.

Browning, S. T. and Shuler, M. L. (2001). Towards the development of a minimal cell model by generalization of a model of *Escherichia coli*: Use of dimensionless rate parameters. *Biotechnology and Bioengineering*, 76(3), 187–192.

Camara, J. E., Crooke, E., and (ed.), N. P. H. (2005). *The Bacterial Chromosome*, chapter 9 - Initiation of Chromosome Replication, pages 177–191. ASM Press.

Castellanos, M., Wilson, D. B., and Shuler, M. L. (2004). A modular minimal cell model: purine and pyrimidine transport and metabolism. *Proceedings of the National Academy of Sciences of the United States of America*, 101(17), 6681–6686. doi:10.1073/pnas.0400962101.

Castuma, C. E., Crooke, E., and Kornberg, A. (1993). Fluid membranes with acidic domains activate DnaA, the initiator protein of replication in *Escherichia coli*. *Journal of Biological Chemistry*, 268(33), 24665–24668.

Chassagnole, C., Noisommit-Rizzi, N., Schmid, J. W., Mauch, K., and Reuss, M. (2002). Dynamic modeling of the central carbon metabolism of *Escherichia coli*. *Biotechnology and Bioengineering*, 79(1), 53–73.

Chiaramello, A. E. and Zyskind, J. W. (1989). Expression of *Escherichia coli* dnaA and mioC genes as a function of growth rate. *Journal of Bacteriology*, 171(8), 4272–4280.

Churchward, G., Estiva, E., and Bremer, H. (1981). Growth rate-dependent control of chromosome-replication initiation in *Escherichia coli*. *Journal of Bacteriology*, 145(3), 1232–1238.

Crooke, E., Castuma, C. E., and Kornberg, A. (1992). The chromosome origin of *Escherichia coli* stabilizes DnaA protein during rejuvenation by phospholipids. *Journal of Biological Chemistry*, 267(24), 16779–16782.

Crooke, E., Thresher, R., Hwang, D. S., Griffith, J., and Kornberg, A. (1993). Replicatively active complexes of DnaA protein and the *Escherichia coli*

chromosomal origin observed in the electron microscope. *Journal of Molecular Biology*, 233(1), 16–24. doi:10.1006/jmbi.1993.1481.

Domach, M. M., Leung, S. K., Cahn, R. E., Cocks, G. G., and Shuler, M. L. (2000). Computer model for glucose-limited growth of a single cell of *Escherichia coli* B/r-A. reprinted from biotechnology and bioengineering, vol. 26, issue 3, pp 203-216 (1984). *Biotechnology and Bioengineering*, 67(6), 827–840.

Donachie, W. D. (1968). Relationship between cell size and time of initiation of DNA replication. *Nature*, 219(5158), 1077–1079.

Donachie, W. D. and Blakely, G. W. (2003). Coupling the initiation of chromosome replication to cell size in *Escherichia coli*. *Current Opinion in Microbiology*, 6(2), 146–150.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*. John Wiley & Sons, Inc.

Garner, J., Durrer, P., Kitchen, J., Brunner, J., and Crooke, E. (1998). Membrane-mediated release of nucleotide from an initiator of chromosomal replication, *Escherichia coli* DnaA, occurs with insertion of a distinct region of the protein into the lipid bilayer. *Journal of Biological Chemistry*, 273(9), 5167–5173.

Gutenkunst, R. N., Atlas, J. C., Casey, F. P., Kuczenski, R. S., Waterfall, J. J., et al. (2007). SloppyCell, `http://sloppycell.sourceforge.net/`.

Hansen, F. G., Atlung, T., Braun, R. E., Wright, A., Hughes, P., et al. (1991a). Initiator (DnaA) protein concentration as a function of growth rate in *Escherichia coli* and *Salmonella typhimurium*. *Journal of Bacteriology*, 173(16), 5194–5199.

Hansen, F. G., Christensen, B. B., and Atlung, T. (1991b). The initiator titration model - computer-simulation of chromosome and minichromosome control. *Research in Microbiology*, 142(2-3), 161–167.

Helmstetter, C. E. (1996). Timing of synthetic activities in the cell cycle. In F. C. Neidhardt, editor, *Escherichia coli and Salmonella, Cellular and Molecular Biology*, volume 2, chapter V-102, pages 1627–1639. ASM Press.

Herrick, J., Kohiyama, M., Atlung, T., and Hansen, F. G. (1996). The initiation mess? *Molecular Microbiology*, 19(4), 659–666.

Hiraga, S., Ichinose, C., Niki, H., and Yamazoe, M. (1998). Cell cycle-dependent duplication and bidirectional migration of SeqA-associated DNA-protein complexes in *E. coli. Molecular Cell*, 1(3), 381–387.

Katayama, T., Fujimitsu, K., and Ogawa, T. (2001). Multiple pathways regulating DnaA function in *Escherichia coli*: distinct roles for DnaA titration by the *datA* locus and the regulatory inactivation of DnaA. *Biochimie*, 83(1), 13–17.

Katayama, T., Kubota, T., Kurokawa, K., Crooke, E., and Sekimizu, K. (1998). The initiator function of DnaA protein is negatively regulated by the sliding clamp of the *E. coli* chromosomal replicase. *Cell*, 94(1), 61–71.

Keseler, I. M., Collado-Vides, J., Gama-Castro, S., Ingraham, J., Paley, S., et al. (2005). EcoCyc: a comprehensive database resource for *Escherichia coli*. *Nucleic Acids Research*, 33(Database issue), D334–D337. doi:10.1093/nar/gki108.

Kholodenko, B. and Westerhoff, H. (2004). *Metabolic engineering in the post-genomic era*. Horizon Bioscience.

Kim, B. G., Good, T. A., Ataai, M. M., and Shuler, M. L. (1987). Growth-behavior and prediction of copy number and retention of ColE1-type plasmids in *Escherichia-coli* under slow growth-conditions. *Annals of the New York Academy of Sciences*, 506, 384–395.

Kim, B. G. and Shuler, M. L. (1990a). Analysis of pBR322 replication kinetics and its dependency on growth-rate. *Biotechnology and Bioengineering*, 36(3), 233–242.

Kim, B. G. and Shuler, M. L. (1990b). A structured, segregated model for genetically modified *Escherichia coli* cells and its use for prediction of plasmid stability. *Biotechnology and Bioengineering*, 36(6), 581–592.

Kitagawa, R., Ozaki, T., Moriya, S., and Ogawa, T. (1998). Negative control of replication initiation by a novel chromosomal locus exhibiting exceptional affinity for *Escherichia coli* DnaA protein. *Genes & Development*, 12(19), 3032–3043.

Kitchen, J. L., Li, Z., and Crooke, E. (1999). Electrostatic interactions during acidic phospholipid reactivation of DnaA protein, the *Escherichia coli* initiator of chromosomal replication. *Biochemistry*, 38(19), 6213–6221. doi:10.1021/bi982733q.

Kurokawa, K., Nishida, S., Emoto, A., Sekimizu, K., and Katayama, T. (1999). Replication cycle-coordinated change of the adenine nucleotide-bound forms of DnaA protein in *Escherichia coli*. *EMBO Journal*, 18(23), 6642–6652. doi:10.1093/emboj/18.23.6642.

Laffend, L. and Shuler, M. L. (1994a). Ribosomal-protein limitations in

Escherichia coli under conditions of high translational activity. *Biotechnology and Bioengineering*, 43(5), 388–398.

Laffend, L. and Shuler, M. L. (1994b). Structured model of genetic-control via the *lac* promoter in *Escherichia coli*. *Biotechnology and Bioengineering*, 43(5), 399–410.

Lee, A. L., Ataai, M. M., and Shuler, M. L. (1984). Double-substrate-limited growth of *Escherichia coli*. *Biotechnology and Bioengineering*, 26(11), 1398–1401.

Mahaffy, J. M. and Zyskind, J. W. (1989). A model for the initiation of replication in *Escherichia coli*. *Journal of Theoretical Biology*, 140(4), 453–477.

Margulies, C. and Kaguni, J. M. (1996). Ordered and sequential binding of DnaA protein to *oriC*, the chromosomal origin of *Escherichia coli*. *Journal of Biological Chemistry*, 271(29), 17035–17040.

McNeil, L. K., Reich, C., Aziz, R. K., Bartels, D., Cohoon, M., et al. (2007). The national microbial pathogen database resource (NMPDR): a genomics platform based on subsystem annotation. *Nucleic Acids Research*, 35(Database issue), D347–D353. doi:10.1093/nar/gkl947.

Messer, W. (2002). The bacterial replication initiator DnaA. DnaA and *oriC*, the bacterial mode to initiate DNA replication. *FEMS Microbiology Reviews*, 26(4), 355–374.

Morgan, J. J., Surovtsev, I. V., and Lindahl, P. A. (2004). A framework for whole-cell mathematical modeling. *Journal of Theoretical Biology*, 231(4), 581–596. doi:10.1016/j.jtbi.2004.07.014.

Nikolaev, E., Atlas, J., and Shuler, M. L. (2006). Computer models of bacterial cells: from generalized coarse-grained to genome-specific modular models. *Journal of Physics: Conference Series*, 46, 322–326.

Nikolaev, E. V., Atlas, J. C., and Shuler, M. L. (2007). Sensitivity and control analysis of periodically forced reaction networks using the Green's function method. *Journal of Theoretical Biology*, 247(3), 442–461. doi:10.1016/j.jtbi.2007.02.013.

Nikolaev, E. V., Burgard, A. P., and Maranas, C. D. (2005). Elucidation and structural analysis of conserved pools for genome-scale metabolic reconstructions. *Biophysical Journal*, 88(1), 37–49. doi:10.1529/biophysj.104.043489.

Overbeek, R., Bartels, D., Vonstein, V., and Meyer, F. (2007). Annotation of bacterial and archaeal genomes: improving accuracy and consistency. *Chemical Reviews*, 107(8), 3431–3447. doi:10.1021/cr068308h.

Palsson, B. O. (2006). *Systems biology : properties of reconstructed networks*. Cambridge University Press.

Schaefer, C. and Messer, W. (1991). DnaA protein/DNA interaction. modulation of the recognition sequence. *Molecular and General Genetics*, 226(1-2), 34–40.

Schaper, S. and Messer, W. (1995). Interaction of the initiator protein DnaA of *Escherichia coli* with its DNA target. *Journal of Biological Chemistry*, 270(29), 17622–17626.

Sekimizu, K. and Kornberg, A. (1988). Cardiolipin activation of dnaA protein, the initiation protein of replication in *Escherichia coli*. *Journal of Biological Chemistry*, 263(15), 7131–7135.

Shu, J. and Shuler, M. L. (1991). Prediction of effects of amino-acid supplementation on growth of *Escherichia coli* B/r. *Biotechnology and Bioengineering*, 37(8), 708–715.

Shuler, M. L. (1999). Single-cell models: promise and limitations. *Journal of Biotechnology*, 71(1-3), 225–228.

Shuler, M. L. (2005). Computer models of bacterial cells to integrate genomic detail with cell physiology. *Proceedings of the KBM International Symposium on Microorganisms and Human Well-Being, June 30-July 2005, Seoul Korea*.

Shuler, M. L. and Dick, C. (1979). A mathematical model for the growth of a single bacterial cell. *Annals of the New York Academy of the Sciences*, 326, 35–55.

Shuler, M. L. and Domach, M. M. (1983). Mathematical-models of the growth of individual cells - tools for testing biochemical-mechanisms. *ACS Symposium Series*, 207, 93–133.

Skarstad, K., Lueder, G., Lurz, R., Speck, C., and Messer, W. (2000). The *Escherichia coli* SeqA protein binds specifically and co-operatively to two sites in hemimethylated and fully methylated *oriC*. *Molecular Microbiology*, 36(6), 1319–1326.

Snoep, J. L., Bruggeman, F., Olivier, B. G., and Westerhoff, H. V. (2006). Towards building the silicon cell: a modular approach. *Biosystems*, 83(2-3), 207–216. doi:10.1016/j.biosystems.2005.07.006.

Speck, C., Weigel, C., and Messer, W. (1999). ATP- and ADP-dnaA protein, a molecular switch in gene regulation. *EMBO Journal*, 18(21), 6169–6176. doi:10.1093/emboj/18.21.6169.

Tomita, Hashimoto, Takahashi, Shimizu, Matsuzaki, et al. (1997). E-CELL: Software environment for whole cell simulation. *Genome Inform Ser Workshop Genome Inform*, 8, 147–155.

Tomita, M. (2001). Whole-cell simulation: a grand challenge of the 21st century. *Trends in Biotechnology*, 19(6), 205–210.

Torheim, N. K. and Skarstad, K. (1999). *Escherichia coli* SeqA protein affects DNA topology and inhibits open complex formation at oriC. *EMBO Journal*, 18(17), 4882–4888. doi:10.1093/emboj/18.17.4882.

W.H., P., B.P., F., and Teukolsky S.A., V. W. (1988). *Numerical Recipes in C*, chapter 6.1, pages 213–216. Cambridge University Press.

Wold, S., Skarstad, K., Steen, H. B., Stokke, T., and Boye, E. (1994). The initiation mass for DNA replication in *Escherichia coli* K-12 is dependent on growth rate. *EMBO Journal*, 13(9), 2097–2102.

Yung, B. Y. and Kornberg, A. (1988). Membrane attachment activates dnaA protein, the initiation protein of chromosome replication in *Escherichia coli*. *Proceedings of the National Academy of Sciences of the United States of America*, 85(19), 7202–7205.

CHAPTER 4

A GENOMICALLY COMPLETE MINIMAL CELL MODEL

4.1 Abstract

A fully functional cell model with explicit genomic information that mimics many details of cellular regulation has been built. This Minimal Cell Model (MCM) allows an engineer to design experiments that probe the cell's behavioral response to environmental and genetic manipulations. It also serves as a platform for the evaluation of candidate minimal gene sets and for the design of more complicated cell models.

In this modeling framework, a cell consists of:

- Compartments
- Species
- Parameters
- Reactions
- Assignment Rules
- Rate Rules
- Algebraic Rules
- Events
- Constraints
- Functions
- Genetic Loci, Genes, and Gene Clusters

With the exception of genes and gene clusters, all of these structures are based on the corresponding structures in the Systems Biology Markup Language (SBML) (Hucka et al., 2003, 2008). Genes and Gene Clusters are data structures that implement instances of the more basic data structures based on SBML. Specifically, creating a new Gene object in the model will cause Species, Reactions, and Rules that correspond to that gene's RNA and protein products to be created automatically.

This chapter explains what each structure is and how it affects the model implementation and simulation. Examples from the MCM are provided. Furthermore, a method for estimating rate and saturation parameters for reaction processes in the cell is described.

The structures listed above are used to define the parts of a cell model based on input criteria. They are not useful for describing the physiology of the cell; i.e., how the structures relate to one another. For example, DNA replication, RNA transcription, and protein translation are all conceptually separate processes in the model, but practically speaking all of their activities are described in the same lists of assignment rules, rate rules, reactions, and other "cell" structures when the cell model is created. Therefore, this chapter also describes the conceptual fragments of the MCM (i.e., the cell's *modules*), including:

Transport - The movement of nutrients into the cell cytoplasm across the plasma membrane via transport reactions.

DNA Replication - The initiation, process, and termination of DNA synthesis.

Transcription - RNA synthesis from a DNA template.

Translation - Protein synthesis from RNA templates.

Demands - The automatic tracking of limiting species in pseudo reactions that consume several reactants (e.g. synthesis of protein).

Geometry - Cell shape as determined by the cell mass and volume.

Gene Set - The genes that are present in the cell, which are determined by the combination of the minimal gene set proposed by Gil et al. (2004) and supplements that have been added to make the gene set physiologically complete.

The overall modeling strategy used here is based on that originally used by Shuler and Dick (1979) and Domach et al. (1984). Those methods were extended more recently (Browning and Shuler, 2001; Browning et al., 2004; Castellanos et al., 2004, 2007; Nikolaev et al., 2005; Atlas et al., 2008). The new model is a system of discontinuous differential algebraic equations which are solved using the SloppyCell (Gutenkunst et al., 2007a) software package. Many significant updates have been made to the original modeling approach to facilitate creating a much larger model than has been attempted previously. Naming conventions related to the model variables and parameters are discussed in Appendix A, and the full model simulation package along with lists of parameters and equations will be made available at the Minimal Cell Model website discussed in Appendix I.

4.2 Introduction

Although an organism's genome, the blueprint of life, encodes all primary information necessary for cellular organization and function (e.g. networks

of interacting biomolecules, regulation, kinetic rate constants, etc.), a more explicit relation of static genomes to dynamic cell physiology and population response is required to take full advantage of thousands of completely annotated genomes (Overbeek et al., 2005; Shuler, 2005). Specifically, a better understanding of how a phenotype evolves from an organism's genome and is affected by dynamic changes in the external environment still represents a significant challenge for modern biology. In this respect, 2D-annotations of complete genomes in terms of accurate stoichiometric reaction networks have recently become available (Palsson 2004) and are now used to provide instant phenotype snapshots of cellular metabolism under fixed external conditions (Palsson, 2006). However, such static snapshots still cannot predict the network's dynamic control, regulation, and systemic response from the collection of functional units (i.e. reactions) and their individual stoichiometries.

A "minimal cell" is a bacterium with the minimum number of genes necessary to grow and divide in some optimally supportive culture environment. The overall goal of this research is to develop a genomically detailed mathematical model of such a cell, which is referred to as a Minimal Cell Model (MCM). The model is constructed and simulated based on the coarse-grained modeling approach developed by the Shuler group (Shuler and Dick, 1979; Domach et al., 1984; Browning and Shuler, 2001; Browning et al., 2004; Castellanos et al., 2004, 2007; Nikolaev et al., 2005; Atlas et al., 2008). This method was originally used to make a coarse-grained model of *Escherichia coli* (Shuler and Dick, 1979; Domach et al., 1984). The MCM's gene set is determined by the combination of the minimal gene set proposed by Gil et al. (2004) and supplements that have been added to make the gene set

physiologically complete (see Section 4.20).

The new model is a system of discontinuous differential algebraic equations (DAEs) that is solved using the SloppyCell software package for Python (Gutenkunst et al., 2007a). The fundamental basis of the modeling approach is that chemical species which have similar dynamics can be aggregated into single model components. Changes in the masses of these components over time are governed by pseudochemical reactions between them. The rates of pseudochemical reactions are based on proposed kinetic relationships that capture the major dependencies of the process being modeled. The largest departure from this method taken in the current research is that the chemical species are significantly more detailed, and many more species are tracked.

In the new modeling framework, a "cell" is composed of compartments, species (i.e. *chemical* species), parameters, reactions, assignment rules, rate rules, algebraic rules, events, constraints, functions, and genes. These structures are defined in detail in Sections 4.4-4.12. Table 4.1 lists the number of each modeling structure included in the MCM.

These structures are each defined in "modules" that describe various physiological processes in the cell (e.g. DNA replication or central metabolism). Each module is defined in a file that has definitions for parameters, species, reactions, etc. related to that module's function. The modules are discussed in detail in Sections 4.13-4.20.

Table 4.1: Model structures used in the Minimal Cell Model. With the exception of genes and gene clusters, all the modeling structures are analogous to their SBML counterparts (Hucka et al., 2003). Rate, saturation, and inhibition parameters are can be set to values from the literature, or estimated using the procedures described in Section 4.7.3. While there are 241 identified coding loci in the model, only 102 are modeled as single genes. The remaining 139 are lumped into groups that have closely coupled function and dynamics. These lumped groups are here named "gene clusters" (Section 4.12).

Model Structure	Count	Examples
Compartments	4	Cytoplasm, cell membrane, whole cell, medium
Chemical Species	408	Glucose-6P, alanine, mRNAs, proteins
Reactions	570	Fructose-6P synthesis, CTP synthesis
Rate Parameters	570	Mass action or Michaelis-Menten rate constants
Saturation Parameters	581	Michaelis-Menten like saturation parameters
Inhibition Parameters	25	Michaelis-Menten like inhibition parameters
Rate Rules	1	Methylation state of chromosome
Algebraic Rules	1	Cell width (CW)
Events	36	DNA replication initiation, cell division
Constraints	408	Each species mass must be > 0
Genes	241	Protein and stable RNA coding regions
Single Coding Genes	102	$dnaB$, pgi, etc.
Gene Clusters	19	$replisome$, etc.
(Genes in Clusters)	139	Ribosomal proteins, $dnaE$, etc.

4.3 Conventions and Assumptions

Making a chemically and genomically detailed model of a cell requires a myriad of assumptions which may not be standard amongst smaller-scale submodels of metabolic processes. In this section the conventions and assumptions that apply to the rest of the equations and modeling structures in this chapter are described. Many of these assumptions are based on those made by (Domach et al., 1984).

The masses calculated in the simulation correspond to the dry weight of the corresponding feature of a real cell. Therefore, the "total mass" of the computer cell corresponds to the "dry weight" of a cell in the lab (i.e. where all water has been removed).

Small inorganic cellular components (e.g. phosphate, magnesium) are available in excess and never limit growth rate or extent. These compounds, therefore, do not need to be explicitly accounted for. In some cases one of these species must be explicitly included in the stoichiometry of a reaction to have the reactants and products be "balanced". Those species are not produced or consumed by the reaction, because it is assumed that regardless of the reaction rates their concentrations are in excess and not rate-limiting.

The cytoplasm is a well-mixed environment. As such, it is assumed that intracellular reactions are not limited by diffusion.

The densities of the cytoplasm and cell membrane are assumed constant. Because the volume of the cell membrane and cytoplasm can vary independently, it is possible for the net density of the cell to vary (in practice it can experience minor fluctuations, but stays relatively constant once steady-state is reached). The volume of a compartment is related to its mass by a constant density.

Cell division results in two identical daughter cells and occurs instantaneously when the septum formation is complete. Therefore, when cell division occurs, the progress of a single daughter cell is tracked.

The cell is in a chemical medium with constant composition, or the cell population is initially low enough that the change in concentration of

nutrients is negligible over the course of the simulation considered. In addition, it is assumed that the reactor medium is well-stirred, so that the cell is constantly exposed to the same concentration of nutrients. The medium contains an excess of all nutrients that the cell needs to survive. Furthermore, all cell waste products are diluted to near zero and cannot be inhibitory.

The cell is in an anaerobic environment. Because the minimal cell has no genes for aerobic respiration (see Section 4.20.4), it is assumed that the environment is anaerobic to ensure the generation of reactive oxidation species (White, 2000, chap. 14). However, without an electron transport chain, the minimal cell may have a reduced ability to generate reactive oxygen species, which could make it slightly aerotolerant.

This model represents an "average" cell, and chemical species with small numbers of molecules can be deterministically rather than stochastically. The Shuler group has presented examples of how to reconcile differences between stochastic and deterministic predictions in bacterial cell models (Browning et al., 2004). All cells in the population have the same composition at the same point in the division cycle.

The cell is spherical in shape. This is based on the lack of significant cytoskeletal proteins that would help the cell maintain another shape. However, the simulation has also been done assuming a rod geometry and the change from one geometry to the other is straightforward.

4.4 Compartments

A compartment is a space where chemical species are located. The MCM currently has four compartments: cytoplasm, cell membrane, cell, and medium. All model volumes are in units of $\mu\mathrm{m}^3$. The volumes of the cellular compartments are calculated using the constant density assumption. Specifically, the densities of the cytoplasm and cell membrane are assumed to be constant, and their values are based on experimental measurements in *E. coli* (Domach et al., 1984). For more detail on the calculations for cell geometry, see Section 4.19. The natural units for volume at the size scale of a bacterial cell are $\mu\mathrm{m}^3$, and that is used in this dissertation. Concentrations in the cell are assigned units of $\frac{\mathrm{pg}}{\mu\mathrm{m}^3}$, where 1 pg $= 1 \times\ 10^{-12}$ g, as those are the natural length and size scales for the MCM. Medium concentrations are referred to in $\frac{\mathrm{g}}{\mathrm{mL}}$. Note that 1 $\frac{\mathrm{pg}}{\mu\mathrm{m}^3}$ = 1 $\frac{\mathrm{g}}{\mathrm{mL}}$, and thus the units used to refer to internal and external concentrations in the MCM are equivalent. The initial volume of each compartment depends on its initial mass, and the calculation of initial mass is discussed in Section 4.5.1.

4.4.1 Cytoplasm (V_C)

The cytoplasm is the compartment where most metabolic reactions take place. It is assumed that the cytoplasm is well-mixed. The volume of this compartment is calculated using the assignment rule in Equation 4.1.

$$V_C = \frac{M_C}{\rho_{cyto}} \tag{4.1}$$

In Equation 4.1, M_C is the (dry) mass of the cytoplasm, and ρ_{cyto} is defined as:

$$\rho_{cyto} = 0.2584 \frac{\text{pg}}{\mu\text{m}^3}$$

The value for ρ_{cyto} is the same as that used by Domach et al. (1984) for modeling *E. coli*. Because M_C is the sum of the masses of all the cytoplasmic species, the volume depends heavily on the initial masses of all the cell components. The initial volume of the cytoplasm in the MCM is set to $4.90 \times 10^{-1}\ \mu\text{m}^3$.

4.4.2 Cell Membrane (V_M)

The volume of the cell membrane is calculated using the assignment rule in Equation 4.2.

$$V_M = \frac{M_4}{\rho_{membrane}} \tag{4.2}$$

In Equation 4.2, M_4 is the mass of the cell membrane, and $\rho_{membrane}$ is defined as:

$$\rho_{membrane} = 0.5526 \frac{\text{pg}}{\mu\text{m}^3}$$

The value for $\rho_{membrane}$ is the same as that used by Domach et al. (1984) for modeling *E. coli*. In that work, Domach also included a factor $prot_{free}$ to account for a protein-free basis of cell-membrane density. Because the MCM explicitly

tracks the masses of lipids and proteins in the cell membrane, the protein free basis for measurements is no longer necessary.

The initial value of M_4, and therefore of V_M, depends on the initial values of the cytoplasmic species. Literature values for cell membrane thickness lie in the 5-10 nm range (Singer and Nicolson, 1972; El-Hag et al., 2006), so the membrane is assumed to have a thickness of 10 nm (0.01 μm). The membrane requires a sufficient mass of phospholipids to fully encompass the volume of the cytoplasm as calculated by Equation 4.1. Taking this into account, the initial volume of the cell membrane compartment is set to 3.11×10^{-2} μm^3.

4.4.3 Cell (V)

The volume of the whole-cell is calculated using the assignment in Equation 4.3. V represents the total cell volume and is the variable that should be compared to experimental measurements of cell volume.

$$V = V_C + V_M \tag{4.3}$$

The initial size of the whole cell, V, in the MCM is set to 5.21×10^{-1} μm^3.

4.4.4 Medium

The medium is the unbounded external environment from which the cell obtains its nutrients. Because the MCM corresponds to a single cell growing in a steady-state environment, it is assumed that all external compounds are

available at constant concentration. The model cell is provided with an excess of all compounds that are necessary for it to grow and divide given its minimal genome. Finally, because the cell does not have the genes necessary for aerobic respiration, it is assumed that the medium is anaerobic.

The 38 compounds present in the medium are listed in Tables D.1 and D.2 in Appendix D. Concentrations proposed for defined media for *Mycoplasma* strain Y (which is similar to *M. mycoides*) for glucose; free bases A, G, and U; some cofactor precursors; and the amino acids were used to define the medium composition in the MCM (Rodwell, 1969).

No suitable reference for the concentration of folic acid, fatty acids, pantothenic acid, or inorganic ions was available, so their initial external concentrations were set to $1 \times 10^{-3} \frac{\text{g}}{\text{mL}}$. Because the external environment is assumed to be constant, changes in the concentrations of external nutrients could be compensated for by changes in the rate constants for transport reactions. Thus, the particular values for the MCM are somewhat arbitrary.

4.5 Chemical Species

A species (i.e., a *chemical species*) is a pool of a particular reacting chemical in the model. There are 408 distinct chemical species in the MCM, and 359 of those are dynamic (i.e. nonconstant). The distribution of species types is presented in Table 4.2. All species inside the cell (e.g. in the cytoplasm or in the cell membrane) are measured in units of mass (pg), while all species in the external medium are measured in units of concentration ($\frac{\text{g}}{\text{mL}}$). Note that the number of proteins given in Table 4.2 is greater than the number of mRNAs because several

protein products have alternate forms. For example, the free cytoplasmic and integral membrane forms of a transporter protein are counted as two distinct species in the MCM.

Because some genes are lumped into gene clusters (see Section 4.12), the number of proteins and mRNAs tracked explicitly in the model is less than the total number of genes in the minimal gene set (see Section 4.20).

4.5.1 Species Initial Conditions

A chemically detailed model of a bacterial cell must have the initial mass of all its chemical species specified. For many chemical species, even average cell cycle values are not known, let alone detailed concentration information as a function of the cell cycle progression. To obtain initial conditions for the MCM, we make use of data for groups of chemical species published for *E. coli* and make assumptions about how these groups are subdivided in the hypothetical cell (Neidhardt, 1996). Because there is no experimental analog for a minimal cell, we propose that using composition data measured in *E. coli* is a valid first-approximation because it will have a similar chemical make-up to other chemoheterotrophic bacteria.

To derive initial values for chemical masses, the following procedure was used (M. Domach, Carnegie Mellon University, personal communication, October 17, 2007):

1. The minimal cell is assumed to have an average dry mass of about 0.2 pg, which is about 75% of the dry weight of *E. coli* (Neidhardt, 1996).

Table 4.2: Distribution of dynamic chemical species defined in the Minimal Cell Model. There are more protein species than mRNA species because some proteins have alternate forms (e.g., if three protein subunits come together to form a protein transporter, the resulting multimeric protein is counted as a separate species). The three bacterial rRNA genes for the 5S, 16S, and 23S bacterial rRNAs are treated as a gene cluster. The three rRNA species tallied here are the immature and mature forms of the rRNA transcripts, which have been lumped as coarse-grained species, as well as the species that tracks rRNA in ribosomes, Rib_{RNA}.

Species Class	**No. Species**	**Details**
Glycolytic	11	Products of glucose catabolism e.g., Fructose-6-P, pyruvate
Pentose Phosphates	7	e.g. ribulose-5P
Amino Acids	20	e.g., alanine, tryptophan
Lipids	7	Cell membrane precursors, e.g., glycerone-3P, palmitoyl CoA
Ribonucleotides	12	e.g., GTP, GDP, GMP
Deoxyribonucleotides	12	e.g., dGTP
mRNAs	100	e.g., $dnaB_{mRNA}$, pgi_{mRNA}, $replisome_{mRNA}$
tRNAs	20	All the tRNA molecules
rRNAs	3	All the rRNA molecules
Amino acid-tRNAs	20	Amino acids bound to their respective tRNAs
DNAs	1	i.e. chromosome(s)
Proteins	123	e.g., DnaB, Pgi, Replisome
Cofactors	23	Products of cofactor metabolism, e.g., thiamine, FAD
Total	359	

2. Data for the average composition of protein, mRNA, tRNA, rRNA, DNA, lipids, and metabolites in *E. coli* was gathered (Neidhardt, 1996). These weight fractions were assumed to be the same for the MCM.

3. Cell age is defined as $age = t/\tau_D$, where t is the time since the last division, and τ_D is the steady state doubling time. A steady-state growth rate μ_g is also defined. The age distribution, $\phi(age)$, for a culture in continuous steady-state growth with a constant τ_D is given by Equation 4.4 (Powell, 1956).

$$\phi(age) = 2\mu_g e^{-\ln(2)\cdot age} \tag{4.4}$$

   To find the average age of a culture (i.e. the 50th percentile), Equation 4.5 is solved for age_{50}.

$$\int_0^{age_{50}} \phi(age)\, da = 0.5 \tag{4.5}$$

   This yields that the average age of a synchronized, exponentially growing cell population (i.e., age_{50}) is approximately 0.415.

4. Assuming the cell is in balanced growth, the population weighted average mass of a chemical species X in the cell will correspond to when the cell is 41.5% of the way through the division cycle. Using Equation 4.7, the initial mass X_0 is calculated from the average mass $\langle X \rangle$.

$$\langle X \rangle = X_0\, e^{(ln(2)\cdot 0.415)} \tag{4.6}$$

$$\langle X \rangle = 1.33\, X_0 \tag{4.7}$$

5. The average mass of each of the protein, mRNA, tRNA, rRNA, and metabolites groups was set to be equal to the mass fraction calculated in step 2 times the total mass selected in step 1. Then, the mass of at the start of the cell cycle was assumed to be the average value divided by 1.33.
6. The initial mass of DNA was set to the mass of one complete chromosome, which was based on the mass of the sequence of the minimal gene set (see Section 4.5.7).
7. The initial mass of membrane lipids was set to be adequate to "envelope" the cytoplasm of the cell (see Section 4.5.5).

The average component masses used to calculate initial conditions are summarized in Table 4.3, and their initial relative magnitudes are shown in Figure 4.1. These proportions agree with the *E. coli* data that they are derived from. Once the component masses were estimated, the masses of individual chemical species were initialized as described in Sections 4.5.2 - 4.5.7. Table E.1 in Appendix E presents initials masses of each chemical species in the MCM. Note that for certain chemical species involved in Demand objects (Section 4.18), the initial condition is shifted by a small amount (<1%) to ensure that one of the chemicals is initially limiting.

This estimate of initial conditions for each chemical species is instrumental in determining the reaction rate constants in the MCM (see Section 4.7.3). Any information regarding precise average values for particular chemicals in a bacterial cell would yield a more authentic representation of cell behavior. The final simulated birth composition is found by letting the cell establish steady-state replication and differs from this initial estimate. The initial estimate has to be sufficiently realistic to yield a stable behavior in the model cell.

Table 4.3: Initial conditions of groups of macromolecules in the Minimal Cell Model. The average masses from *E. coli* are based on values reported in (Neidhardt, 1996). The average mass in the MCM is calculated by assuming that each component accounts for the same mass percentage in *E. coli* and the minimal cell, but that the total average mass of the minimal cell is 0.2 pg. Note that the actual average value of DNA used in the MCM is based on its genome sequence, not on the data from *E. coli* presented in this table. In the current model the mass of the chromosome is $M_{CHR} \sim 3.77 \times 10^{-4}$ pg. Initial values for the start of the cell cycle were calculated as described in Section 4.5.1.

Component	Avg. mass in *E. coli* (pg)	Avg. mass in MCM (pg)
Protein	1.56×10^{-1}	1.20×10^{-1}
rRNA	4.77×10^{-2}	3.68×10^{-2}
tRNA	6.33×10^{-3}	6.33×10^{-3}
mRNA	2.10×10^{-3}	1.62×10^{-3}
DNA	9.00×10^{-3}	6.95×10^{-3}
Lipid	2.60×10^{-2}	2.01×10^{-2}
Metabolites	1.00×10^{-2}	7.72×10^{-3}

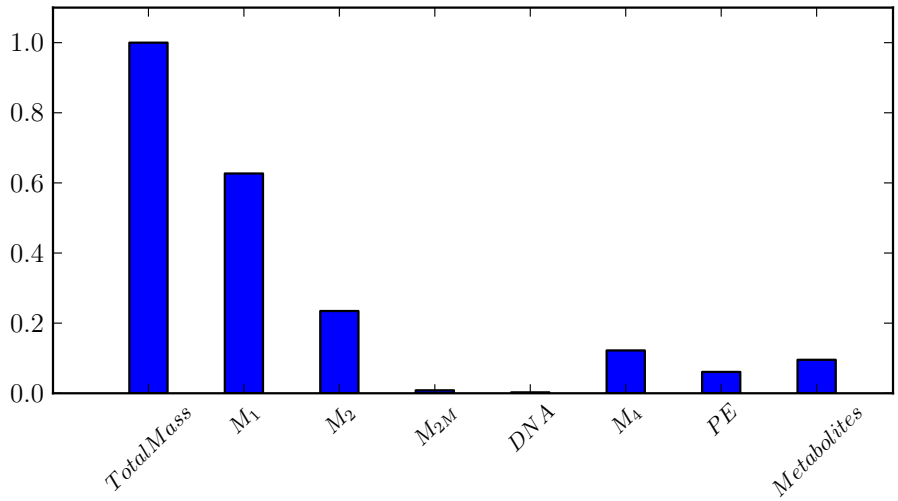


Figure 4.1: Relative initial masses of lumped species groups in the Minimal Cell Model. M_1 is protein, M_2 is total RNA, M_{2M} is mRNA, M_3 is DNA, M_4 is cell membrane (protein and lipid), PE is membrane lipids, $Metabolites$ are all the precursor molecules including nucleotides, amino acids, and sugars.

4.5.2 mRNA and Protein

The minimal cell initially has 1.22×10^{-3} pg of mRNA per cell. The mass of each mRNA species was calculated by taking the desired total initial mass of mRNA and dividing it up evenly amongst the mRNA species for all protein coding genes in the cell. For gene clusters producing a single coarse-grained mRNA species, the initial mRNA amount was weighted by the numbers of genes represented in the cluster. See Section 4.12 for more information on gene clusters.

Initial masses for each protein were calculated similarly, with the exception that membrane and ribosomal proteins were initiated before "free" proteins. The membrane protein content was based on the experimental observation that proteins account for 50%-80% of the weight of membranes in mollicutes (Korn, 1969; Razin, 1975). A 50/50 split of lipid and protein in the membrane is assumed because the minimal cell will have fewer proteins than a naturally occurring cell. The total initial mass of protein in the cell was set to 9.03×10^{-2} pg.

4.5.3 tRNA

The minimal cell initially has 4.75×10^{-3} pg of tRNA per cell. In the MCM, tRNA can exist as a free species or in a bound species with its corresponding amino acid. To set the initial conditions for tRNAs, it was assumed that each free and bound species had the same initial mass of tRNA (subtracting out the mass of amino acid attached to the tRNA).

4.5.4 rRNA

The minimal cell initially has 2.76×10^{-2} pg of rRNA per cell including 23S rRNA, 16S rRNA, and 5S rRNA. However, it is unclear how much of this rRNA exists in free form separate from any ribosomes in nature. As a starting point, it was assumed that the free immature and mature rRNA species each have six times the initial mass as each free mRNA. The remainder of the initial mass of rRNA is added to ribosomes.

4.5.5 Lipids

The initial amount of lipids present in the cell membrane of the minimal cell is calculated from physical constraints on the shape and mass of the cell rather than from the literature data. Specifically, given the density of the cell membrane, the cell requires a minimum amount of lipid material to sufficiently "envelope" the cytoplasm.

The membrane is assumed to be 0.01 μm thick (Singer and Nicolson, 1972). Using this thickness the volume, and therefore mass, of membrane is calculated based on the approximate mass of the cytoplasm. For example, for a spherical cell the volume of the membrane is approximately expressed as,

$$V_M \approx \frac{4}{3}\pi \cdot \left(\frac{CW}{2} + d_M\right)^3 - V_C \tag{4.8}$$

$$M_{4init} = \rho_{env} * V_M \tag{4.9}$$

where M_{4init} is the initial mass of the membrane, V_M is its volume, CW is the cell width, V_C is the volume of the cytoplasm, and d_M is the thickness of the membrane.

The membrane is a 50%-80% mixture of proteins and lipids (Korn, 1969; Razin, 1975). For the MCM the initial membrane protein mass is set to half of the membrane's mass because it is expected that a minimal cell has fewer membrane proteins than a traditional cell. Therefore, the initial mass of PE in the membrane is assumed to be one-half of the calculated initial membrane mass M_{4init}. The remaining membrane mass is divided amongst the membrane transport proteins (Section 4.5.2). Septum material is generated as part of the cell division process, and its mass (referred to as $sept$ in the MCM) is initially set to zero.

4.5.6 Metabolites

Metabolites are defined as all the precursors of macromolecules and cofactors of biosynthetic reactions in bacterial metabolism. The minimal cell initially has 5.79×10^{-3} pg of metabolites per cell, and these species are initialized such that the sum of all metabolite masses is equal to the desired initial sum. This results in an unusual initial distribution of metabolites that changes drastically once a simulation commences based on the demand for those metabolites in the cell. Metabolism is related to protein and mRNA synthesis as well as genome sequence, and those nonlinear effects lead to the model compounds achieving a new steady state once the simulation begins.

For chemicals that enter the cell via diffusion an extra low initial mass was

selected to guarantee an inward facing concentration gradient. Specifically, the concentration of each diffusing chemical in the cytoplasm was set to be one tenth of the concentration in the medium.

4.5.7 Genome

The MCM's gene set is determined by the combination of the minimal gene set proposed by Gil et al. (2004) and supplements that have been added to make the gene set physiologically complete (see Section 4.20). To determine the sequence of each genetic locus in the minimal gene set, a Python script has been written that automatically downloads the required nucleotide sequences, along with the corresponding protein sequences, from the Kyoto Encyclopedia of Genes and Genomes (KEGG) website at `http://www.genome.jp/kegg/` (Kanehisa and Goto, 2000). This script will be included at the supplementary website described in Appendix I. The KEGG database allows one to search for gene sequences from a variety of organisms, and the organisms used in the search are described in Section 4.20.7.

The genome sequence and chromosome mass are calculated from the sequence of the minimal genome. The initial state of the model is assumed to be just after a successful round of DNA replication and cell division, such that the initial mass of DNA should be set to the mass of a single chromosome. Because the minimal genome is drastically different than the genome of *E. coli*, in particular in its abbreviated length, the MCM will have a significantly lower initial and average mass of DNA than *E. coli* (Table 4.3).

4.6 Parameters

A parameter is a named quantity in the model that is not a species or a compartment. The implementation of parameters used here is based on that described in the SBML documentation (Hucka et al., 2008). There are constant parameters (e.g. rate constants), and nonconstant parameters (e.g. cell width). Most of the constant parameters in the MCM are created automatically when reactions are defined. Every reaction has a single rate constant and one saturation parameter for every saturation chemical associated with the reaction (these correspond to activation or inhibition terms). The nonconstant parameters are set continuously by assignment rules, rate rules, or algebraic rules (Section 4.8), or discontinuously by event assignments (Section 4.9). Nonconstant parameters include gene dosages, which are set by assignment rules as described in Section 4.16.

4.7 Reactions

A reaction governs the conversion of one set of species (reactants) into another (products). There are 570 reactions in the MCM, and a small subset of those are discussed here.

Reactions are defined by their stoichiometry and rate law. Stoichiometry is based on the mass of products that will be produced when a given amount of reactants are consumed. Rate laws in the model are written in terms of the production rate of one of the product species (i.e., the calculated rate is in units of $mass/time$ of product produced per cell). All stoichiometries in

the model are mass based. It is possible to input reaction stoichiometry on a molar basis, but this is automatically converted to a mass basis once the reaction object is defined. This allows us to write differential equations more easily for species in a cell with changing volume. For reactions that consume ATP or other phosphate donors, coupled phosphate donor consumption reactions are introduced as part of the reaction stoichiometry. Again, these reactions are mass based, so the consumption of ATP is always in terms of the mass of ATP consumed per unit mass of product formed. If the product of a reaction is specified, then the stoichiometry of that reaction is normalized according to the mass of product produced.

4.7.1 Inputs for a Reaction Object

Rate laws for each reaction are automatically constructed based on a number of inputs. Specifically, one can specify a reaction's:

- stoichiometry
- rate constant
- saturation term(s)
- external saturation term(s)
- inhibition term(s)
- rate multiplier(s)
- flag(s)
- enzyme

The rate constants are necessary for all reactions, even if they are set to 1.0. The modeling framework tolerates unknown rate constants as well, which can be automatically estimated using the method described in Section 4.7.3. The units of the rate constants vary depending on the form of the rate law, but they are usually in $\frac{mass_{product}}{time}$ or $\frac{mass_{product}}{time \cdot mass_{enzyme}}$ units.

Saturation terms are Michaelis-Menten type terms of the form:

$$\left(\frac{X}{X + V \cdot K_S}\right) \tag{4.10}$$

where X is the mass of chemical species X, V is the volume of the compartment containing species X, and K_S is the saturation constant in $\frac{mass}{volume}$ units. In the MCM, both reactants and cofactors can have saturation effects on a reaction rate. Allowing cofactors to participate in saturation terms ensures that if there is no cofactor, the reaction rate eventually drops to zero.

Similarly, external saturation terms are terms that account for the saturating effect of an extracellular species on a reaction rate. These terms, which usually only occur in transport reactions, have the form:

$$\left(\frac{X_{ext}}{X_{ext} + Ks_{ext}}\right) \tag{4.11}$$

where X_{ext} is the concentration of species X outside the cell in $\frac{mass}{volume}$ units, and Ks_{ext} is the external saturation constant in $\frac{mass}{volume}$ units.

Reactions can also have inhibition terms of the form:

$$\left(\frac{K_i}{K_i + \frac{X}{V}} \right) \tag{4.12}$$

where X is the mass of chemical species X, V is the volume of the compartment where the reaction is taking place, and K_i is the inhibition constant $\frac{mass}{volume}$ units.

Multiplier and flag terms have similar effects on the rate law of a reaction. They are simply terms that get multiplied into the reaction rate. Enzyme masses are used as multiplier terms to encompass the effect of a changing net catalytic activity on the reaction rate. For example, in the generic rate law presented in Equation 4.13, E is the mass of an enzyme that catalyzes the reaction, and the term E operates as a rate multiplier.

$$rate = v_m \cdot E \cdot \left(\frac{X}{X + V \cdot K_S} \right) \tag{4.13}$$

Flags work in the same way as multipliers (i.e., their values are multiplied into the rate law). However, flags can be zero-valued and are used to simulate times when the reaction is shut off. Because they can be set to 0, their effect is ignored when calculating rate constants.

Reactions that have a catalyzing enzyme are handled by adding a multiplier to the rate law encompassing the effect that enzyme has on the rate. Pseudo-reactions that depend on multiple enzymes (e.g., DNA synthesis) use "pseudo-enzymes" whose masses represent the sum of all the enzymes involved in that process. Sections 4.7.4-4.7.6 show examples of reactions governing synthesis of fructose-6-P, the dATP deoxyribonucleotide, and DNA.

4.7.2 Determination of Saturation Parameters

Saturation constants for activation terms in saturation-type rate laws were estimated by applying a general rule of thumb that postulates that a reasonable value for an unknown saturation constant is one twenty-fifth of its normal intracellular concentration (NIC) (Domach et al., 1984). Similarly, inhibition constants for inhibition terms in rate laws are estimated by applying a heuristic that the constant will be equal to 10 times that chemicals NIC. In the MCM, the NIC is set to the predicted average concentration of each chemical species. This rule has been applied in previous models (Shu and Shuler, 1991; Domach et al., 1984).

4.7.3 Rate Constant Estimation

Developing a model of this scale is complicated by lack of kinetic information for most of the proposed reactions. One could estimate the rate constants for the reactions in the model one at a time, but as the number of reactions increases it becomes more difficult to select rate constants that allow simulation of a viable (i.e. repeatedly dividing) cell. At the same time, parameter analysis research has revealed that in many biological models, the specific values of parameters are not as critical as their ratios to one another (Browning and Shuler, 2001; Brown and Sethna, 2003; Gutenkunst et al., 2007b,c). For that reason, a method to quickly estimate rate constants for coarse-grained models of single cells growing at steady-state has been developed. The goal of developing this procedure is not to calculate rate constants for the enzymes that could be used in another context. Rather, the goal is to rapidly obtain a reasonable set of

parameters that can be used to help test the plausibility candidate minimal gene sets. This method is based on the following assumption:

Assumption: In a single cell growing and repeatedly dividing at steady-state, each chemical species' mass will double in the time that it takes for the cell to divide, τ_d.

This assumption is certainly true in an exponentially growing population of bacterial cells experiencing balanced growth, and applying the assumption to the single-celled model allows us to calculate rate constants for the reactions in the model. We begin by using the doubling assumption for species X_i (i.e. $X_i(t_d) = 2 \cdot X_i(0)$) to write:

$$\int_0^{t_d} \frac{dX_i}{dt}\, dt = X_i(t_d) - X_i(0) = X_i(0) \tag{4.14}$$

The rate $\frac{dX_i}{dt}$ is not constant, but for most species the mass X_i will increase monotonically until it doubles in a nearly linear fashion. We can take advantage of this to calculate a set of approximate rate constants that are likely to result in a cell model that will achieve a stable cell division cycle. Specifically, it is assumed that the rate of production of a species X_i is linear in the rate constants v_j, and that the nonlinear portions of the rate laws are known functions of $\mathbf{X}$, $f_j(\mathbf{X})$, which is approximated over the course of the cell doubling time (Equation 4.15). Furthermore, it is assumed that each species creates a constraint on some of the rate constants as in Equation 4.16.

$$\frac{dX_i}{dt} = \sum_{j=0}^{N_R} v_j \cdot \alpha_{i,j} \cdot f_j(\mathbf{X}) \tag{4.15}$$

$$\sum_{j=0}^{N_R} v_j \cdot \alpha_{i,j} \cdot f_j(\mathbf{X}) \geq ss_i \cdot \frac{X_i(0)}{t_d} \tag{4.16}$$

Specifically, Equation 4.16 says that the sums of all the reaction rates acting on species i are constrained to being greater than $X_i(0)$, the mass of species i at time 0, divided by the desired doubling time. In Equation 4.16, a scaling factor, ss_i, is introduced for certain chemical species that are under-produced using this procedure. A value of ss_i greater than 1 directs the rate constant estimation algorithm to attempt to produce more than two times as much of species i before the end of the cell cycle. For example, DNA (M_3), which doubles its mass in a fraction of the cell division cycle, will be under-produced using the estimation procedure outlined here. To ensure sufficient production of DNA, ATP, and CoA by the cell, the target accumulation rates of these species were adjusted to obtain rate constants that resulted in a repeating cell cycle. Specifically, ss_i was set to 2.0 for DNA, and to 4.0 for ATP and CoA.

While the assumption of linearity is not true (because $f_j(\mathbf{X})$ is nonlinear), by applying this assumption to the initial conditions for the MCM, linear constraints on the rate constants for the model are obtained. This results in a system of constraint equations on all the rate constants in the model, which can be expressed as a matrix A. The objective function:

$$f_{opt} = \sum_{i=1}^{N_R} v_i \tag{4.17}$$

where N_R is the number of reactions, and v_i is the rate constant for rate constant i, is introduced to frame the problem as a Linear Programming (LP) problem with constraints A and objective function f_{opt}, which is minimized. The space

of possible rate constant choices is a many dimensional space and there can be infinitely many sets of constants that would satisfy the given constraints. The objective function is minimized because the constraints placed on the reaction rate constants (doubling all chemical species masses) tend to force the system to have higher rate constants. To balance these constraints and estimate reasonably sized rate constants, their sum is minimized. The LP system is solved using the Python lpsolve package (Berkelaar et al., 2010). A wrapper class for lpsolve is included with the MCM code.

It is possible that there is some prior information available about the value of a rate constant for a particular reaction in the model. In those cases, upper and lower bounds on the rate constant are incorporated into the LP method. This ability is useful in cases where, for example, a single transport enzyme operates on several substrates. While it is not necessarily true that the transport rate constant will be the same for all substrates, they are likely constrained to similar ranges. Thus, to obtain an appropriate set of constants the rate constants for transporters with multiple substrates are constrained to being equal.

The rate constant estimation procedure described above allows us to obtain a reasonable set of parameters that can be used to help test the plausibility of candidate minimal gene sets. The absolute values of the parameters selected is in some sense arbitrary for an MCM (Browning and Shuler, 2001). It is of note that the parameters estimated here will have different values if the initial conditions of the MCM are altered. Furthermore, two similar reactions (e.g., two protein synthesis reactions) may yield different rate constants if their products are consumed differently in the cell (e.g., a cytosolic protein and a protein that is transported to the membrane). This is acceptable for the base MCM as long

as reasonable values are achieved for reactions with known biochemistry.

To ensure that the rate constant calculation procedure calculates reasonable values, the values calculated are compared to values from bacteria that have been measured, as in Table 4.4. For each comparison, the rate constant from the MCM is converted to a specific activity by recognizing that,

$$activity = \frac{v_m}{E} \tag{4.18}$$

where E is the mass of the enzyme corresponding to v_m.

4.7.4 Reaction f6P$_S$

f6P$_S$ is the reaction catalyzed by glucose-6P isomerase (Pgi), and it converts glucose-6P (g6P) into fructose-6P (f6P). Because this is an isomerization reaction, the mass of the reactant and product are the same, and each has a stoichiometric coefficient of $\pm$ 1 (Equation 4.19). The rate law for this reaction (Equation 4.20) consists of the rate constant $v_{f6P\text{-}S}$ ($\frac{\text{pg f6P}}{\text{h·pg Pgi}}$), a saturation term for the reactant with saturation constant $Ks_{f6P\text{-}S\text{-}g6P}$ ($\frac{\text{pg}}{\mu\text{m}^3}$), and a multiplier for the mass of Pgi enzyme per cell (pg).

$$1 \cdot \text{g6P} \xrightarrow{\text{Pgi}} 1 \cdot \text{f6P} \tag{4.19}$$

$$\left(\frac{df6P}{dt}\right)_{f6P_S} = v_{f6P\text{-}S} \cdot \frac{g6P}{(g6P + Ks_{f6P\text{-}S\text{-}g6P} \cdot V_C)} \cdot Pgi \tag{4.20}$$

Table 4.4: Comparison of reaction rate constants estimated for the MCM to values for fermentative bacteria (der Werf et al., 1997). The specific activity is shown in units of $\frac{\text{nmol P}}{\text{min·mg M}_1}$, where P product formed and M_1 is the mass of protein in the system. To calculate specific activity rates from the reaction rate constants in the MCM, the rate constant (in units of $\frac{\text{pg P}}{\text{h·pgE}}$) was first converted into $\frac{\text{nmol P}}{\text{min·pg E}}$, and then the ratio of E to M_1 (masses of E and M_1, respectively) in the cell was used to adjust the activity to a per protein total basis.

Enzyme	**Specific Activity** $\left(\frac{\text{nmol P}}{\text{min·mg protein}}\right)$		
	***Actinobacillus* sp. 130Z**	***E. coli* K-12**	**MCM**
6-Phosphofructokinase (PfkA)	900	890	272
fructose-1,6-diphosphate aldolase (FbaA)	1,800	960	272
Glyceraldehyde-3-phosphate dehydrogenase (GapA)	810	850	575
Enolase (Eno)	8,300	3,500	576
Pyruvate kinase (PykA)	1,000	1,400	240
Lactate dehydrogenase (Ldh)	9	560	546

4.7.5 Reaction dATP$_{\mathrm{S}}$

dATP$_{\mathrm{S}}$ is the reaction catalyzed by adenylate kinase (Adk), which catalyzes the interconversion of adenine nucleotides. The stoichiometric coefficients for the reaction have been normalized in terms of the product, dATP (Equation 4.21). Note that the mass-based stoichiometric coefficients on each side of the reaction have the same sum. Thus, the chemical reaction is balanced in mass. In the rate law for dATP$_{\mathrm{S}}$ (Equation 4.22), $v_{dATP\text{-}S}$ is the maximum rate of the reaction ($\frac{\text{pg dATP}}{\text{h}\cdot\text{pg Ndk}}$), $Ks_{dATP\text{-}S\text{-}dADP}$ is the saturation constant describing the effect of $dADP$ on the rate ($\frac{\text{pg}}{\mu\text{m}^3}$), $Ks_{dATP\text{-}S\text{-}ATP}$ is the saturation constant describing the effect of ATP on the rate ($\frac{\text{pg}}{\mu\text{m}^3}$), and Ndk is the mass of Ndk per cell (pg).

$$0.837 \cdot \text{dADP} + 1.03 \cdot \text{ATP} \xrightarrow{\text{Ndk}} 0.87 \cdot \text{ADP} + 1 \cdot \text{dATP} \tag{4.21}$$

$$\left(\frac{d(dATP)}{dt}\right)_{dATP_S} = v_{dATP\text{-}S} \cdot \frac{dADP}{(dADP + Ks_{dATP\text{-}S\text{-}dADP} \cdot V_C)} \cdot \frac{ATP}{(ATP + Ks_{dATP\text{-}S\text{-}ATP} \cdot V_C)} \cdot Ndk \tag{4.22}$$

4.7.6 Reaction M$_{3-\mathrm{S}}$

M$_{3-\mathrm{S}}$ is a pseudoreaction for the synthesis of DNA from dNTP precursors. Note that the stoichiometric coefficients of the reactants and products have the same sum (Equation 4.23), so the reaction is balanced in mass. As a pseudoreaction, it is catalyzed by a sum total of proteins which is simply called "Replisome". The Replisome consists of gene products from the *dnaE*, *dnaN*, *dnaQ*, *dnaX*,

holA, *holB*, *gyrA*, *gyrB*, *lig*, and *ssb* genes. $DNAp_{min}$ is the mass of the most "in-demand" dNTP species at a given time, and ATP is the mass of cellular ATP. It is assumed that the DNA replication process depends on the energy of the cell (Domach and Shuler, 1984), and that dependence is represented by the ATP saturation term. The rate law is given in Equation 4.24, where $NTOT$ is a parameter that tracks the number of actively replicating forks on the chromosome. μ_3 is the maximum rate of the reaction per replication fork ($\frac{\text{pg M}_3}{\text{h·pg Replisome}\cdot Fork}$), $Ks_{M3\text{-}S\text{-}DNAp\text{-}min}$ is the saturation constant describing the effect of dNTPs on the rate ($\frac{\text{pg}}{\mu\text{m}^3}$), $Ks_{M3\text{-}S\text{-}A2}$ is the saturation constant describing the effect of glycolytic compounds on the rate ($\frac{\text{pg}}{\mu\text{m}^3}$), and $Replisome$ is the mass of the gene-cluster product corresponding to genes involved in DNA synthesis.

$$\left.\begin{array}{rl} & 1.05\times10^{-1}\cdot\text{ATP}\\ +\ & 3.01\times10^{-1}\cdot\text{dATP}\\ +\ & 1.93\times10^{-1}\cdot\text{dCTP}\\ +\ & 2.10\times10^{-1}\cdot\text{dGTP}\\ +\ & 2.96\times10^{-1}\cdot\text{dTTP}\\ +\ & 3.72\times10^{-3}\cdot\text{H2O}\end{array}\right\} \xrightarrow{\text{Replisome}} \left\{\begin{array}{rl} & 1.00\cdot\text{DNA}\\ +\ & 8.78\times10^{-2}\cdot\text{ADP}\\ +\ & 3.56\times10^{-4}\cdot\text{AMP}\\ +\ & 2.01\times10^{-2}\cdot\text{Pi}\\ +\ & 1.83\times10^{-4}\cdot\text{PPi}\end{array}\right. \tag{4.23}$$

$$\left(\frac{dM3}{dt}\right)_{M3_S} = mu_3\cdot\frac{ATP}{(ATP+Ks_{M3\text{-}S\text{-}ATP}\cdot V_C)} \cdot\frac{DNAp_{min}}{(DNAp_{min}+Ks_{M3\text{-}S\text{-}DNAp\text{-}min}\cdot V_C)} \cdot Replisome\cdot NTOT \tag{4.24}$$

The stoichiometry of DNA synthesis is calculated when the cell model is

Table 4.5: ATP consumption related to chromosome synthesis.

Process	ATP Consumption	Reference
Histone HupA	1 ATP per HupA molecule	assumption
Helicase DnaB	1 ATP per DnaB molecule	assumption based on (White, 2000)
Gyrase	2 ATP per 10 bp	assumption based on (White, 2000)
Ligase	1 ATP per 1000 bp	assumption

defined, based on the sequence of DNA in the cell. The model does not link the consumption of each dNTP precursor to the position of the replication fork. Rather, the consumption of dNTPs is executed on an average basis. Similarly, the ATP consumption is calculated on an average basis, as described in Table 4.5.

4.8 Rules

Rules provide a means to control the values of variables in a model. The implementation of rules used in the MCM is based on that proposed by SBML (Hucka et al., 2008). There are three subclasses of rules based on the following three functional forms (where X is a variable, f is some function, $\mathbf{V}$ is a vector of variables that does not include X, and $\mathbf{W}$ is a vector of variables that may include X) (Hucka et al., 2008):

Assignment Rules - Those rules where the left hand side is the value of the variable set by that rule, i.e. $X = \mathrm{f}(\mathbf{V})$.

Rate Rules - Those rules where the left hand side is the value of the rate of change of the variable set by that rule, i.e. $\frac{\mathrm{d}X}{\mathrm{d}t} = \mathrm{f(V)}$.

Algebraic Rules - Those rules where the left hand side is zero, i.e. $0 = \mathrm{f}(\mathbf{W})$

The three classes of rules are described in detail in Sections 4.8.1 - 4.8.3. More rigorous definitions of rules are available in the SBML specification (Hucka et al., 2008).

4.8.1 Assignment Rules

Assignment rules are used to express equations that set the value of a variable, and an implementation of assignment rules based on that described in the SBML documentation is used (Hucka et al., 2008). These rules are usually used as a means of calculation convenience, and they are used extensively in the MCM to track the masses of "lumped species" such as amino acids (P_1) or the total mass of the cell. For example, the assignment rule for the total mass of amino acids (P_1) is,

$$\begin{aligned} P_1 &= Val + Tyr + Gln + Gly + Glu + Ala + His \\ &+ Pro + Ser + Phe + Asn + Thr + Cys + Leu \\ &+ Ile + Asp + Trp + Lys + Arg + Met \end{aligned} \tag{4.25}$$

Some of the lumped species defined in the MCM are presented with their general definitions in Table 4.6.

Table 4.6: Some lumped species defined by assignment rules in the Minimal Cell Model. Each species is actually a parameter whose value is set to the sum of all the chemical species in that lumped species. There are other lumped species in the model omitted here for brevity.

Lumped Species	**Definition**
A_2	All compounds involved in glycolysis.
P_1	All 20 amino acids.
P_2	Ribonucleotides
P_3	Deoxyribonucleotides
P_4	Cell membrane precursors
PPP	All compounds involved in the pentose phosphate pathway
$cofactors$	All cofactors included for cofactor metabolism
M_1	All protein species.
M_2	All RNA species.
M_{2M}	All mRNA species.
$tRNA$	All tRNA species.
M_4	Total cell membrane mass (protein and lipid).
M_C	Total mass of all cytoplasmic species.
$TotalMass$	Total mass of all chemical species.

4.8.2 Rate Rules

A rate rule expresses the rate of change of a particular variable, and an implementation of rate rules based on SBML is used here (Hucka et al., 2008). Variables that are set by rate rules may not appear in chemical reactions, and therefore in this model no chemical species trajectories are set directly by rate

rules. Rate rules are, therefore, used to expresses the rates of change for particular cell parameters. The only rate rule used in the current model is Equation 4.26.

$$\frac{dMethState}{dt} = MethRate \tag{4.26}$$

$MethState$ is a parameter that describes how methylated the chromosome is (on a scale from 0 to 1). The $MethRate$ in Equation 4.26 refers to the rate at which methyl groups are transferred from S-adenosylmethionine (sam) to DNA by the MraW enzyme, forming S-Adenosyl-L-homocysteine (sahs) as a by-product (Equation 4.27).

$$MethRate = v_{meth} \cdot sahs_S \tag{4.27}$$

In previous models produced in the Shuler group, rate rules were also used to track the rate at which septum was formed in the cell division process (Domach et al., 1984; Browning and Shuler, 2001). In the current model, septum material is an explicitly modeled chemical species, so no artificial rate rule is necessary. Equation 4.26 gives the rate of change of the methylation state of DNA, which affects the ability of the cell to initiate DNA replication.

4.8.3 Algebraic Rules

An algebraic rule describes a constraint on a model variable in relation to other model variables, and an implementation of algebraic rules based on SBML is used here (Hucka et al., 2008).

The MCM uses a single algebraic rule (Equation 4.28) to constrain the width of the cell (CW), as described in Section 4.19.

$$0 = V - (V_{cellbody} + V_{endcaps} + V_{septum}) \tag{4.28}$$

In Equation 4.28, $V_{cellbody}$, $V_{endcaps}$, and V_{septum} are defined by assignment rules based on the width, length, and surface area of the cell. The variable that is adjusted to make the geometric constraint true is the cell width, CW.

4.9 Events

Events describe instantaneous, discontinuous changes in the state of the model, and an implementation of events based on SBML is used here (Hucka et al., 2008). Because they cause discrete changes in the cell structure or behavior that occur instantaneously when the cell reaches some predefined condition, events require special mathematical treatment during a simulation. For example, the 'initiation of DNA replication' event occurs when a threshold number of DnaA molecules is bound to the DNA $OriC$.

In the MCM, an event could, for example, describe instantaneous changes in the masses of the chemical species in the cell (i.e. at cell division). There are a total of 36 events in the base model. The names and trigger functions for all 36 events are presented in Appendix F. Here, we present as examples a generic event, as well as the "DNA Initiation" and "DNA Termination" events from the MCM. The entire list of events is summarized in Table F.1. The specifics of the DNA replication events will be discussed in Section 4.15.

4.9.1 Generic Event Example

Imagine an event where the concentration of a metabolite ($elicitor$) activates the synthesis of a species in a secondary metabolic pathway. When the concentration of the elicitor is above a threshold, the event is triggered, i.e.

$$\textbf{Trigger: } \frac{elicitor}{V} > thr_e$$

The event will occur when the concentration of the elicitor ($\frac{elicitor}{V}$) is greater than the threshold, thr_e. Once the trigger function's value changes from false to true, the event "fires", and the cell responds by executing a number of event assignments. In the case of the elicitor, one might expect a number of reaction pathways to be activated or augmented. For example,

Event Assignments:

$$v_x \rightarrow 1e^6$$

$$flag_e \rightarrow 1$$

where v_x is some reaction rate constant that is increased to a new level by the presence of the elicitor, and $flag_e$ represents that some other physiological process has been activated.

4.9.2 DNA Initiation

DNA Initiation is the start of chromosome synthesis. The trigger function for DNA Initiation is:

Trigger: $(DnaG_{boundto\text{-}Ori} \geq init_{threshold}) \wedge (flag_{meth} == 1)$

In short, the replication process is triggered when the mass of $DnaG$ bound to the origin of replication (Ori) exceeds threshold $init_{threshold}$. The trigger function for DNA replication initiation is explained in more detail in Section 4.15.1.

There are 21 event assignments associated with DNA replication initiation. Below, 11 examples are presented.

Sample DNA Initiation Event Assignments:

$$DnaG_{boundto\text{-}Ori} \rightarrow 0$$

$$DnaB_{boundto\text{-}Ori} \rightarrow 0$$

$$HupA_{boundto\text{-}Ori} \rightarrow 0$$

$$DnaG \rightarrow DnaG + DnaG_{boundto\text{-}Ori} \cdot Ori_{GD}$$

$$DnaB \rightarrow DnaB + DnaG_{boundto\text{-}Ori} \cdot Ori_{GD}$$

$$HupA \rightarrow HupA + HupA_{boundto\text{-}Ori} \cdot Ori_{GD}$$

$$flag_{meth} \rightarrow 0$$

$$MethState \rightarrow 0$$

$$flag_{repl} \rightarrow 1$$

$$M3_{init} \rightarrow DNA$$

$$t_{DNA\text{-}init} \rightarrow time$$

After DNA replication commences, it is assumed that the proteins bound to the Ori are rapidly forced off by the opening of the chromosome replication fork. Thus, $DnaG_{boundto\text{-}Ori}$, $DnaB_{boundto\text{-}Ori}$, and $HupA_{boundto\text{-}Ori}$ are set to zero by this event. Those proteins are each added back into the cytoplasmic pools, and the free pools of $DnaG$, $DnaB$, and $HupA$ are updated to reflect the change.

Some event assignments reflect changes in the cell's state. For example, setting $flag_{meth}$ and $MethState$ to 0 resets the methylation state of the chromosome, and setting $flag_{repl}$ to 1 tells the model that DNA replication is now active, so that the DNA synthesis reaction is activated. Other event assignments are updates of bookkeeping parameters. For example, $M3_{init}$ is recorded to monitor the initiation mass of the cell, and $t_{DNA\text{-}init}$ tracks the time of replication initiation. The full list of event assignments will be available on the website described in Appendix I

4.9.3 DNA Termination

The simple trigger function for DNA replication termination becomes true when the replication fork reaches the terminus of replication (see Section 4.15.3).

Trigger: $(ForkPos_0 \geq 1.0)$

After DNA replication ends, 11 variables are updated. For example, C_{period} tells how long the chromosome replication took. The full list of event assignments associated with DNA Termination will be available on the website described in Appendix I

4.10 Model Failure and Constraints

Model failure in the minimal cell corresponds to cell death, but cell death can occur for a variety of reasons. Each reason is considered to be analogous to a particular type of cell failure in biology.

Constraints are a mechanism for specifying the conditions under which the model simulation is invalid, and an implementation of constraints based on that described in the SBML documentation is used here (Hucka et al., 2008). In the MCM, constraints that specify that no species can have a negative mass are introduced. There is one constraint for each chemical species. Constraints are implemented in SloppyCell as events that cause a Mass Constraint Violation exception to be raised in Python that lets the user know that an invalid model condition has been encountered (Gutenkunst et al., 2007a).

For a cell to be viable it must have a stable cell division cycle. If the model just continually grows without dividing, a parameter set has been selected for which the model's trajectory through state space does not intercept the cell division event surface (i.e., the Poincaré map for the system is not approaching a fixed point (Nikolaev et al., 2005)). This error may also lead to an OverflowError in the Python simulation as the concentration of a particular species will increase indefinitely if the cell never divides.

A Singularity Error occurs in the solution when there is a non-invertible matrix encountered in the integration. This points to an inadequacy or inconsistency in the model solution or parameter set.

A Zero Division Error occurs when the cell has need of a nutrient that is not present. This is similar to a Constraint Violation, because no species should be able to be consumed in the cell to the point where its concentration is zero (although it may become infinitesimally small).

4.11 Functions

Function definitions associate an identifier with a function such that the identifier can be used to call the specified function anywhere else in the model (Hucka et al., 2008). The only function defined in the current MCM is the Heavy Function (HF), which is convenient for determining the gene dosage for each gene in the model. More detail on the HF function definition is provided in Section 4.16, and simple examples of functions are presented in the SBML documentation (Hucka et al., 2008).

4.12 Genetic Loci, Genes, and Gene Clusters

A genetic locus is a location on the computer chromosome that may code for protein or RNA products and may bind various species.

For a single genetic locus, the user can specify its:

- chromosomal positions (multiple copies are allowed)
- binders (chemical species that bind to the locus)
- DNA sequence
- number of genes (for gene clusters)
- source organism
- functional category and subcategory

A gene is a genetic locus that codes for either a protein or RNA product. Gene clusters are groups of genes that perform related functions and are

adjacent to each other on the computer chromosome. For the purposes of the model simulation gene clusters are treated as single genes. All products from a gene cluster are assumed to be subject to the same global regulation mechanisms. Genes within a cluster are positioned together on the chromosome. These could be groups of operons or regulons, but practically speaking the impact on the cell model is that the protein products of all the genes in the cluster are treated as a single lumped species.

A computer chromosome is automatically constructed from the genes in the MCM's minimal gene set. There is conflicting evidence regarding the conservation of gene order in bacteria (Mushegian and Koonin, 1996a; Dandekar et al., 1998; Tamames et al., 2001; Tamames, 2001). Mushegian and Koonin (1996a) found that gene order is not generally conserved in distantly related bacteria. Dandekar et al. (1998) reported that genes whose orders are conserved are more likely to have physically interacting products. One line of research found that bacterial shape was determined by the order of genes involved in cell division rather than by their presence or absence (Tamames et al., 2001), and that gene order is conserved in closely related species (Tamames, 2001). The prevailing evidence, however, has suggested that gene order is not conserved across long evolutionary distances in bacterial species (Mushegian and Koonin, 1996a; Tamames, 2001), and thus the genes are ordered arbitrarily in this first release of the full MCM. It is proposed to implement a more rigorous scheme for gene ordering in future work (Section 6.2).

For *coding* genes (i.e. those that code for protein or RNA products), one can specify their:

- mRNA and protein initial concentrations
- protein sequence (in the case where the sequence is different from what you would predict based on the DNA sequence).
- factors influencing the mRNA and protein synthesis or degradation rates

Each gene has a variable "gene dosage" (GD) that tells how many copies of that gene exist in the cell at a given time. Note that the gene dosage for a "gene cluster" is multiplied by the number of genes in that cluster. This accounts for the fact that for a cluster with n genes, the corresponding RNA transcript species represents n transcript products.

4.12.1 Binders

Some genetic loci are capable of binding proteins or other molecules in the cell. The extent to which a particular locus is bound can affect cell physiology. For example, the initiation of DNA replication depends on the binding of proteins to the origin of replication. Binding molecules could also be used to implement transcriptional level control of gene expression, though this is not used in the current model. The unbinding rates for binders are written as first-order rate laws. However, because the concentration of binding molecules is quite low, the unbinding rate is insignificant compared to the binding rate in the default condition of the model.

4.12.2 Gene Products

RNA products include tRNA, rRNA, and mRNA species. The synthesis reactions for these products are described in Section 4.16.

Protein product synthesis rates are described in Section 4.17. A user can optionally specify an alternate form for protein products when initiating a coding gene. This is useful for cases where a protein can be converted into another species through metabolic reactions or physiological species. For example, proteins translocated into the membrane to act as integral membrane transporters have a "free" form and a "membrane" form.

4.13 Transport

Our strategy for simulating a minimal cell depends largely on importing the building blocks of macromolecules from the external environment. The model treats glucose, fatty acids, free nucleoside bases, all 20 amino acids, cofactor precursors, and some inorganic ions as external species at a constant concentration that must be transported into the cell. For each transporter a gene, or genes, that correspond to the transporter are included. The protein products of these genes must be incorporated into the membrane to be catalytically active.

Transport proteins are synthesized in the same manner as all other proteins in the model. A Michaelis-Menten expression is used to describe integration of transporters into the membrane because it is assumed to be an enzyme catalyzed process controlled by chaperon proteins coded for by the $prot_{transloc}$ gene cluster described in Section 4.13.1. Transporters affect transport reactions

as multipliers to their rates. Rate equations for nutrient import and waste export are written either as Michaelis-Menten like equations or as simple diffusion equations. Rate constants are estimated for transport equations using the procedure outlined in Section 4.7.3. It is of note that enzymes with multiple substrates are constrained to have similar transport rate constants for each of their substrates.

This model includes four types of transport: primary active transport coupled to phosphate transfer, active transport coupled to H^+ or Na^+ import (symport), facilitated diffusion promoted by transport proteins, and passive diffusion driven by a concentration gradient.

Active transport is the transport of species across the membrane against a concentration gradient. Primary active transport uses chemical energy (such as ATP) to provide energy for the movement against a concentration gradient. To model active transport we write transport reactions that include the simultaneous consumption of the appropriate phosphate donor. For example, consider the stoichiometry and rate of the alanine (Ala) transport reaction (Equations 4.29 and 4.30).

$$0.202 \cdot \mathrm{H2O} + 5.7 \cdot \mathrm{ATP} + 1 \cdot \mathrm{Ala_{ext}} \xrightarrow{\mathrm{T_{Bgt}}} 4.8 \cdot \mathrm{ADP} + 1 \cdot \mathrm{Ala} + 1.1 \cdot \mathrm{Pi} \quad (4.29)$$

$$\left(\frac{dAla}{dt}\right)_{R_{Ala}} = v_{R\text{-}Ala} \cdot \frac{Ala_{ext}}{(Ala_{ext} + Ks_{R\text{-}Ala\text{-}Ala\text{-}ext})} \cdot \frac{ATP}{(ATP + Ks_{R\text{-}Ala\text{-}ATP} \cdot V_C)} \tag{4.30}$$

$$\cdot \frac{Ki_{R\text{-}Ala\text{-}Ala}}{\left(Ki_{R\text{-}Ala\text{-}Ala} + \frac{Ala}{V_C}\right)} \cdot T_{Bgt} \cdot Ki_{R\text{-}Ala}$$

$Ki_{R\text{-}Ala}$ is defined as a product of Michaelis-Menten inhibition terms (Equation 4.31).

$$Ki_{R\text{-}Ala} = \prod_{i=1}^{N} \frac{Ki_{R\text{-}Ala\text{-}i}}{Ki_{R\text{-}Ala\text{-}i} + Inhib_i} \tag{4.31}$$

In Equation 4.30, $\left(\frac{dAla}{dt}\right)_{R_{Ala}}$ is the rate of the transport reaction for alanine catalyzed by the Bgt transporter, $v_{R\text{-}Ala}$ is a transport rate constant ($\frac{\text{pg Ala}}{\text{h·pg T}_{\text{Bgt}}}$), Ala_{ext} is the external concentration of alanine ($\frac{\text{g}}{\text{mL}}$), Ala is the cytoplasmic mass of alanine, ATP is the cytoplasmic mass of ATP (pg), T_{Bgt} is the membrane mass of Bgt transporter protein, $Ks_{R\text{-}Ala\text{-}Ala\text{-}ext}$ is a saturation constant describing the activating effect of external alanine on the uptake rate, ($\frac{\text{g}}{\text{mL}}$), $Ks_{R\text{-}Ala\text{-}ATP}$ is a saturation constant describing the activating effect of cytoplasmic ATP on the uptake rate ($\frac{\text{pg}}{\mu\text{m}^3}$), and V_C is the cytoplasmic volume (μm^3). $Ki_{R\text{-}Ala}$ is a dimensionless inhibition term describing the competitive inhibition of alternate substrates for the Bgt transport system and defined in Equation 4.31 (see Section 4.13.1). The inhibition constants in Equation 4.31 all have units of $\frac{\text{pg}}{\mu\text{m}^3}$, and there are N such constants.

A second type of active transport used in the MCM is symport, which is when a substrate is transported into the cell against its concentration gradient while another small molecule or ion is transported into the cell with its

concentration gradient. For example, the transport of aromatic amino acids into the cell is facilitated by AroP, an H^+ symporter. The stoichiometry and rate laws for symport reactions are analogous to Equations 4.30 and 4.31. The model does not track the cellular or extracellular concentration of H^+, so the strength of the proton motive force is not explicitly known. However, the Gil et al. (2004) gene set includes a set of genes from ATP synthase to maintaining the proton gradient (Gil et al., 2004). These genes are included as a gene cluster, but the cell also boots the proton motive force by coupling proton export to lactate export (see Section 4.13.8).

Facilitated diffusion is driven using rate laws of the form in Equation 4.30. The only difference between this active transport rate and a facilitated diffusion rate is that facilitated diffusion does not have the ATP dependence. The stoichiometry for such a reaction would not include any ATP consumption either. Currently, the only nutrient imported into the MCM via facilitated diffusion is K^+.

Simple diffusion is the spontaneous transport of a species across the membrane in the same direction as its concentration gradient. For example, thiamine is transported into the cell via a diffusion reaction (Equations 4.32 and 4.33).

$$1 \cdot \text{thiamine}_{\text{ext}} \longrightarrow 1 \cdot \text{thiamine} \tag{4.32}$$

$$\left(\frac{d(thiamine)}{dt}\right)_{T_{thiamine}} = v_{R\text{-}thiamine} \cdot \left(thiamine_{ext} - \frac{thiamine}{V_C}\right) \cdot SA \tag{4.33}$$

where $v_{R\text{-}thiamine}$ is a transport rate constant related to the diffusion coefficient and the cell membrane thickness ($\frac{\mu\text{m}}{\text{h}}$), $thiamine_{ext}$ is the external concentration of thiamine ($\frac{\text{g}}{\text{mL}}$), $\frac{thiamine}{V_C}$ is the cytoplasmic concentration of thiamine ($\frac{\text{g}}{\text{mL}}$), and SA is the cell surface area. As expected, the model has diffusion transport rates that are significantly lower than the active transport rates. However, if experimental evidence shows that the cofactor precursors which are assumed to enter the cell through simple diffusion cannot enter quickly enough to maintain a growing cell, it may be necessary to introduce new genes whose products can facilitate their diffusion.

4.13.1 Transporter Function

Some protein transporters operate on multiple substrates. In these cases, there is competition between the substrates for the enzyme. A product of multiple Michaelis-Menten competitive inhibition terms is used to account for the effect of multiple substrates. Each transport rate law has one inhibition term for each alternative substrate. For example, a transporter that carries four substrates will have three external inhibition multipliers for each of its transport rate laws. The inhibition constant for each inhibition term is assumed to be 15x the default external concentration for each inhibitory nutrient. It should be noted, however, that because the external environment of the cell is constant, that any change in inhibition term can be exactly compensated for by an adjustment in the rate constant for the corresponding reaction. Therefore, the absolute values of the inhibition constants used in the development of this base model are not critical as long as they correspond to reasonably achievable nutrient concentrations.

In *E. coli*, transport proteins share the same pathway for localization as some excreted proteins (Murphy and Beckwith, 1996). Gil et al. (2004) recommend a set of five genes for protein translocation and secretion. These five genes are included as a single gene cluster, $prot_{\text{transloc}}$, whose product catalyzes the incorporation of transport proteins into the membrane. This cluster includes the genes *ffh*, *ftsY*, *secA*, *secE*, and *secY*. The process is exhibited in Figure C.1.

4.13.2 Transport in the Minimal Gene Set

Gil et al. (2004) propose including only two transport systems in the minimal cell. They included a phosphotransferase (PTS) system for active-transport of carbohydrates, and a transporter for inorganic phosphate (Pi) to provide phosphate for metabolic reactions. The implementation of the PTS system is discussed in Section 4.13.3, and the function of the PitA transport system is explained in Section 4.13.7. Gil et al. (2004) also propose that the products of the *hpt* and *upp* genes catalyze a simultaneous transport and activation of external free bases, but they do not label these genes as "transport" specific (Section 4.13.4). Similarly, they propose that the uptake and activation of fatty acids is performed by the product of the *fadD* gene (Section 4.13.5).

We have elected to include 18 additional transporters in the MCM for amino acid and ion transport (Sections 4.13.6-4.13.7). The Gil et al. (2004) gene set proposes that lactate will be the end product of the fermentation of glucose, but it proposes no means to remove that lactate from the cell. To prevent excessive lactate accumulation in the cell a lactate transporter has been included in the MCM (Section 4.13.8). In total, the MCM has 23 genes that code for

proteins whose primary purpose is transport. Of these, 19 are new additions compared to the Gil et al. (2004) minimal gene set. Finally, precursors of cofactor biosynthesis are allowed to enter the cell via diffusion (Section 4.13.9) as proposed by Gil et al. (2004).

4.13.3 Glucose Transport

The phosphotransferase (PTS) system imports and phosphorylates glucose at the expense of phosphoenolpyruvate (PEP). The PTS transporter included is a glucose-specific PTS system coded by *ptsG*, *ptsH*, and *ptsI*. Gil et al. (2004) found that all the components of the PTS were present in all the reduced genome bacteria they considered except for *W. glossinidia*. PTS plays a crucial role in the MCM because it provides a means for the cell to obtain both energy and carbon for metabolism. The PTS system in the MCM is feedback inhibited by glucose-6P and activated by the presence of PEP.

4.13.4 Nucleotide Precursor Transport

To synthesize nucleotides and then later RNA and DNA, the cell needs to be able to import free bases. Analysis of the *M. genitalium* genome has not identified any genes that code explicitly for free base transporters (Mushegian and Koonin, 1996b). Castellanos et al. (2004) chose to include a single gene product for free base transport, but did not specify which gene coded for it in their final gene tally. The *nupG* gene in *E. coli* is responsible for nucleoside import, but those nucleosides must be altered to get transformed into NMPs. Gil et al. (2004)

propose that free bases diffuse through the minimal cell's simple membrane, but they claim it is also possible that their transport and incorporation is coupled to the *hpt* and *upp* reactions (Hochstadt-Ozer and Stadtman, 1971). Here it is assumed that the free bases A, G, and U are transported into the cell in reactions catalyzed by membrane bound Hpt and Upp.

4.13.5 Fatty Acid Transport

Fatty acid biosynthesis pathways were incomplete in most of the genomes studied in (Gil et al., 2004). Based on the minimal gene set they proposed Gil et al. (2004), it is assumed that the transport of fatty acids into the cell is coupled to the action of acyl-CoA synthase (EC 6.2.1.3), which is encoded by *fadD* (Gil et al., 2004; Schmelter et al., 2004).

4.13.6 Amino Acid Transport

Gil et al. (2004) proposed that amino acids diffuse into the cell through its less highly structured cell membrane (Gil et al., 2004). A functional minimal cell would probably need protein transporters to ensure that amino acids are delivered at a rate capable of sustaining growth. Therefore, the Gil et al. (2004) minimal gene set is supplemented with 14 genes related to transport of amino acids. The amino acid transporters are summarized in Table 4.7. All amino acid transporters used in the MCM require energy either directly from ATP or indirectly in cotransport.

Table 4.7: Amino acid transporters in the Minimal Cell Model. ABC: ATP Binding Cassette permease. SDF: Sodium:dicarboxylate (SDF) symporter. The transporters are drawn from *Escherichia coli*, *Synechocystis* sp., *Rhodobacter sphaeroides*, and *Bacillus subtilis*.

Transporter	**Genes**	**Transport Family**	**Amino Acids**	**Reference**
AroP	*aroP*	H^+ Symport	Trp, Tyr, Phe	(Sarsero et al., 1991; Burkovski and Krämer, 2002)
Bgt	*bgtA*, *bgtB*	ABC	Ala, Gln, Gly, Leu, Pro, Ser	(Quintero et al., 2001)
BztD	*bztD*	ABC	Asn, Glu, Gln, Asp	(Zheng and Haselkorn, 1996)
LivF	*livF*	ABC	Val, Leu, Ile	(Riley et al., 2006)
MetT	*metI*, *metN*, *metQ*	ABC	Met	(Merlin et al., 2002)
Nat	*natA*, *natB*, *natC*, *natD*	ABC	Arg, His, Lys	(Quintero et al., 2001)
SstT	*sstT*	SDF Symport	Ser, Thr	(Burkovski and Krämer, 2002)
TcyP	*tcyP*	SDF Symport	Cys	(Burguire et al., 2004)

4.13.7 Inorganic Ion Transport

Transporters for K^+, Mg^{2+}, Mn^{2+}, Na^+, and inorganic phosphate (Pi) have been included in the MCM. Of these ions, the Gil et al. (2004) gene set only provides a transporter for Pi. Inorganic cations are necessary for three reasons in prokaryotes (Silver, 1996):

1. The cells require a high intracellular osmolarity to maintain turgor pressure.
2. Cations are reusable cofactors in some enzymes.
3. Metalloenzymes use the cations as stably-bound permanent components.

The minimal gene set proposed by Gil et al. (2004) does not include protein transporters for any inorganic cations. Instead, they propose that a minimal cell would obtain these ions from the environment via diffusion in a manner akin to the "free-diffusing cell" proposed by Luisi et al. (2002). It would probably be difficult for the cell to obtain the ions it requires from the environment by diffusion alone. Therefore, in the MCM, one transporter for each of the categories listed above is included, as well as an Na^+:H^+ antiporter to export Na^+ that accumulates from the import of serine and threonine. Note that the MCM does not track the masses of individual ions. The mathematical effect of having the proposed transporters included is to make sure that the model accounts for the metabolic energy and precursors required in their synthesis. An estimate of the cation uptake rate is also made for use in calculating the energy requirements for the transport process.

Potassium (K^+) is found in the cytoplasm of all organisms at high concentration (Silver, 1996), and this high concentration helps the cell maintain its turgor pressure. *E. coli* has at least six enzyme systems related to the uptake or export of K^+ (Silver, 1996). From these the Kup system has been selected, which is a single-protein, low-affinity uptake system controlled by chemiosmotic force (Silver, 1996). The low-affinity system was selected because it contains only a single protein, and the idealized environment of the minimal cell will have a sufficiently high level of K^+ to provide a concentration gradient for transport.

Magnesium (Mg^{2+}) is the second most abundant intracellular cation after potassium (Silver, 1996), but in contrast to K^+ its role is primarily as a cofactor of some enzyme catalyzed reactions. For Mg^{2+} uptake the MgtA system of *E. coli* was selected (Silver, 1996). This is an ATP binding protein and therefore an estimate of Mg^{2+} uptake rate is made to calculate how much ATP is required for the transport system.

Manganese (Mn^{2+}) is an example of a metal that is stably bound to some metalloenzymes (Silver, 1996). The energy source for the transport of (Mn^{2+}) in *E. coli* is the proton motive force, and that consumption is accounted for in the MCM's submodel for energy metabolism. The gene that codes for this transporter is *mntH* (Kehres and Maguire, 2003; Courville et al., 2004).

Sodium export ensures that sodium imported from the uptake of serine and threonine does not accumulate to toxic levels. The MCM includes the *nhaB* gene to code for an Na^+:H^+ antiporter.

The cell requires a phosphate transporter to replenish the inorganic phosphate (Pi) consumed in the synthesis of phosphorylated nucleotides and their precursors. Gil et al. (2004) include the gene *pitA* in their proposed gene set. PitA is a Pi transporter powered by the proton motive force.

K^+, Mg^{2+}, and $PO_4{}^{3-}$ are all reported as a necessary components of growth media for *Mycoplasma mycoides Y* (Miles, 1992), which supports inclusion of their transporters in this model. Although this model is significantly more detailed than previous versions of the MCM, it does not yet have the resolution to monitor the concentrations of ions. Introducing ion concentration tracking would allow a better understanding of energy metabolism processes. However, it is still necessary to estimate the rate at which the ions are transported to calculate their approximate impact on energy metabolism.

Mn^{2+} and Pi import, as well as Na^+ export, are powered by the membrane energization, or proton motive force, of the cell. It is assumed that each divalent ion will required the symport of two H^+, while each the monovalent sodium ion will require one H^+ carried in antiport. To figure out how many protons need to be imported for these processes, it is necessary to know approximately how many ions are being transported.

For Mn^{2+} and Pi import, the estimate made in the Shuler group's *E. coli* model is followed and it is assumed that all the inorganic ions account for 5% of the cell's mass (Domach, 1983). If the minimal cell weighs approximately 1 pg and it is assumed that the ions have equal representation, then each ion will account for 0.01 pg. The net transport over a cell cycle is approximated with the expression,

$$\int_{t=0}^{t=t_d} v_{ion} \cdot K_{satext} \cdot T_{ion}\, dt = ion_0 \tag{4.34}$$

where t is time, t_d is the doubling time for the cell, v_{ion} is the maximum transport rate for the ion, K_{satext} is the constant saturation term for the external concentration of the ion, T_{ion} is the mass of the transport enzyme in the cell membrane as a function of time, and ion_0 is the initial mass of the ion in the cell. The initial mass for each transport enzyme in the MCM is about 7.5×10^{-4} pg. If it is assumed that the enzyme's mass will double according to the exponential rate law $T_0 \cdot 2^t$, then the integration in Equation 4.34 can be performed and estimate for the unknown rate constant v_{ion} can be obtained. This yields,

$$v_{ion} = \frac{ion_0}{K_{satext} \cdot \int_{t=0}^{t=t_d} T_{ion}\, dt} = \frac{ion_0 \cdot ln(2)}{K_{satext} \cdot T_0} \tag{4.35}$$

Equation 4.35 provides an estimate for the ion uptake rate constant which is then used to estimate how many H^+ ions need to be exported by the ATP synthase to maintain the proton motive force. This rate constant estimate is used for each of the ion transporters in the MCM.

For Na^+ the transport burden is indirectly coupled to the amount of serine and threonine imported into the cell. It is assumed that each molecule of serine or threonine is symported into the cell with a single Na^+. To prevent an accumulation of Na^+ in the cell, it is assumed that the vast majority of sodium must be exported. Therefore, the number of protons that must be exported to account for sodium export is directly calculable from the rates of serine and threonine uptake.

While the MCM does not maintain a detailed balance of inorganic ions, these calculations ensure that the model accounts for the physiological energy burden of ion uptake.

4.13.8 Lactate Transport

The final reaction of glycolysis based on the minimal gene set is the pyruvate dehydrogenase reaction, which converts pyruvate into lactate. Because all the ATP production in the minimal cell comes from substrate level phosphorylation, glycolysis runs at a high throughput and a large amount of lactate will be produced. The follow-up work to Gil et al. (2004), considers lactate to be a "sink" chemical (Gabaldón et al., 2007), but they do not specify the mechanism by which its concentration is maintained. It is proposed here that the minimal cell will require a mechanism for lactate efflux.

Lactate efflux from bacteria is common in fermentative bacteria, but specific proteins involved in the process are not very well studied (Konings et al., 1995). Research suggests that lactate can be released via symport with 1-2 H^+ (Konings et al., 1994; Konings, 2002). This expulsion of H^+ ions is in accordance with Gil's suggestion that the ATPase acts as a proton pump to maintain the membrane polarization with respect to H^+ (see Section 4.20.4). Therefore, the lactate permease (*lctP*), which has been shown to have lactate export activity in *B. subtilis* (Ramos et al., 2000; Chai et al., 2009), is included. This proton symport system translocates 1-2 H^+ for every molecule of lactate. Because one H^+ is produced per lactate during lactic acid fermentation, the proton symport must export two H^+ per lactate to yield a net change in the proton motive force.

However, in an acidic environment, the lactate export will only result in one H^+ per lactate molecule being translocated into the medium. Therefore, at low pH values lactate excretion does not generate a membrane potential (Konings et al., 1994). The ATP synthase included in Gil et al. (2004) will run in reverse to maintain the membrane potential.

4.13.9 Diffusive Transport

Gil et al. (2004) propose that all of the nutrients necessary for cofactor biosynthesis diffuse into the cell. This includes thiamine, riboflavin, nicotinamide, folic acid, and pantothenic acid. The general rate equation for diffusion is presented in Equation 4.36, where R_i is the rate of diffusion, P_i is the permeability per unit thickness of the membrane to species i, SA is the surface area of the cell, and C_{out} and C_{in} are the concentrations of the species outside and inside the cell, respectively.

$$R_i = P_i \cdot SA \cdot (C_{out} - C_{in}) \tag{4.36}$$

4.14 Metabolic Reactions

One module is dedicated to defining the reactions of metabolism, including glycolysis, the pentose phosphate pathway, nucleotide biosynthesis, membrane lipid biosynthesis, and cofactor biosynthesis. The overall metabolism of the MCM is presented in Figure 4.2.

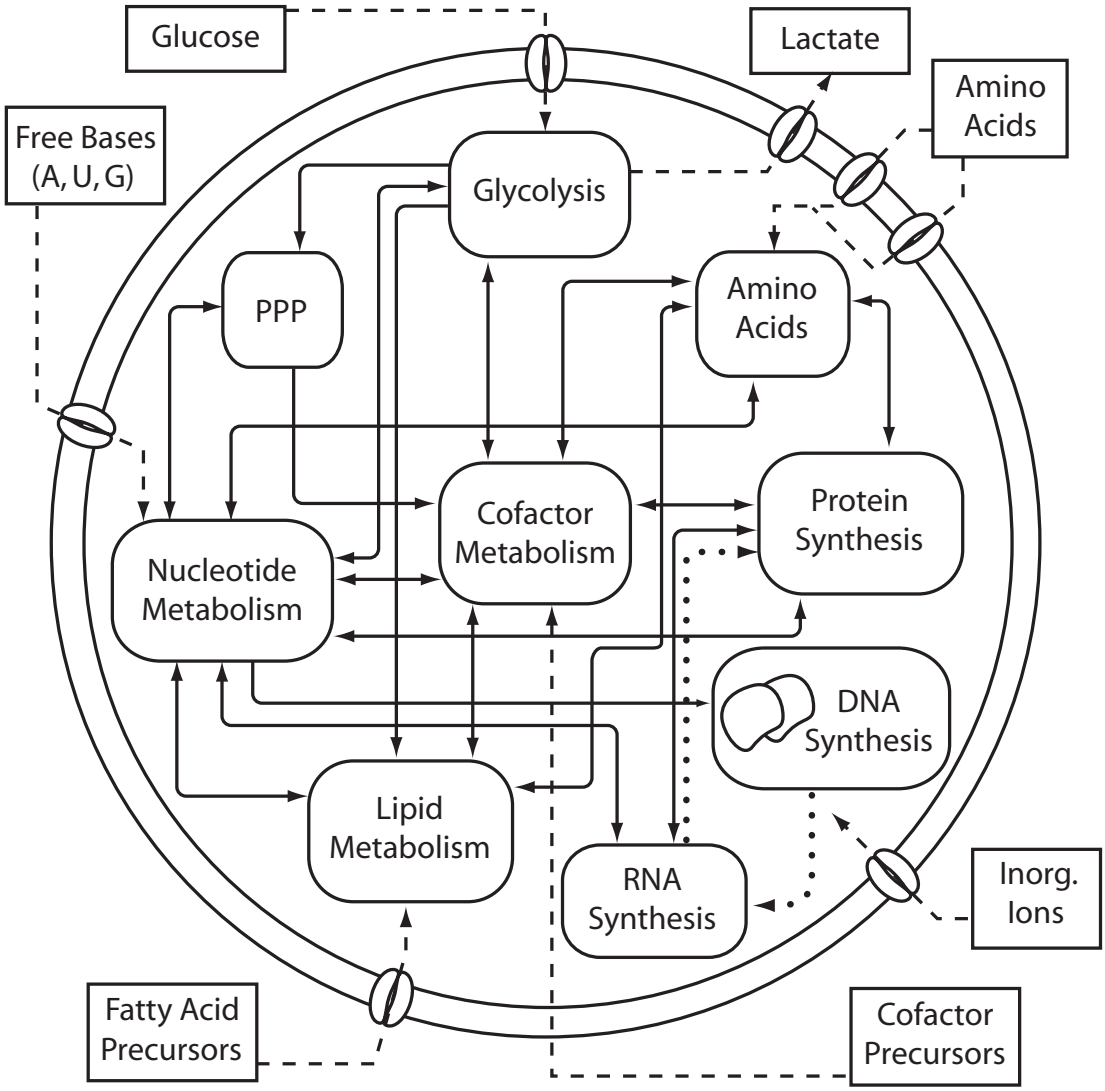


Figure 4.2: Overview of metabolic processes included in the Minimal Cell Model. Solid lines represent flow of mass within the cell. Dashed lines represent transport processes. Boxes within the cell membrane are subsets of metabolism described by the MCM. External nutrients for the MCM include glucose, amino acids, inorganic ions, cofactor precursors, fatty acid precursors, and free bases. PPP is the Pentose Phosphate Pathway. Details of reactions in each box are displayed in Appendix C in Figures C.2-C.8.

4.14.1 Glycolysis

The minimal cell depends heavily on the bacterial glycolytic pathway. Through glycolysis, the cell can synthesize ATP from substrate level phosphorylation as well as make precursors necessary for the pentose phosphate pathway and for lipid metabolism. All the major enzymes of glycolysis are included in this pathway (see Figure C.2). It is of note that there is a strong feedback for glucose uptake from one of the end-products of glycolysis, PEP.

4.14.2 Pentose Phosphate Pathway

The pentose phosphate pathway takes fructose and phosphoglyceraldehyde from glycolysis and uses it to synthesize 5-carbon sugars which are precursors for nucleotide biosynthesis. Gil et al. (2004) include *rpe* (ribulose-phosphate 3-epimerase), *rpiA* (Ribose 5-phosphate isomerase), and *tkt* (transketolase) in their original minimal gene set (Gil et al., 2004). They suggest in an update that *glpX* (sedoheptulose-1,7-bisphosphatase) is also necessary (Gabaldón et al., 2007). All of these genes have been included in the MCM. The pathway is depicted in Figure C.3.

4.14.3 Lipid Metabolism

The proposed minimal gene set contains seven genes dedicated to lipid biosynthesis. This is in contrast to an earlier MCM that proposed a lipid synthesis module with only five genes (Castellanos et al., 2007). In the analysis of Castellanos et al. (2007), it is assumed that glycerol-3-phosphate and

fatty acids are each transported into the cell by membrane transport proteins. However, no genes are explicitly named to accomplish this function, accounting for the discrepancy in gene counts (Castellanos et al., 2007). Gil et al. (2004) explicitly includes a gene that transports fatty acids into the cell, as well as a gene to synthesize glycerol-3-phosphate from glycolytic intermediates.

In accordance with Gil et al. (2004) it is assumed that the minimal cell has a lipid bilayer made only of phosphatidylethanolamine (PE) with embedded membrane proteins. They conclude that the fatty acid precursors necessary for membrane biosynthesis can be obtained from the environment and activated in a single step by acyl-CoA synthase (*fadD*). In a follow-up paper, Gabaldón et al. (2007) imply that the specific acyl-CoA used is palmitoyl CoA (pal), and this assumption is followed here.

Gil et al. (2004) include *plsB* and *plsC* for converting the activated fatty acids into phosphatidate (PA), as well as *cdsA* to convert the PA into CDP-diglyceride. Castellanos et al. (2007) also included these genes. The remaining genes differ in (Castellanos et al., 2007) and (Gil et al., 2004) because the former assumed that a minimal cell's membrane would be composed of phosphatidylglycerol while the later assumed it would be phosphatidylethanolamine (PE). The assumption of Gil et al. (2004) is followed and PE is used as the membrane phospholipid.

The full lipid biosynthesis pathway is depicted in Figure C.4.

4.14.4 Nucleotide Metabolism

The minimal cell synthesizes all ribonucleotides (and deoxyribonucleotides) from the A, G, and U free bases, in combination with the 5-phosphoribosyl diphosphate sugar (PRPP). Phosphate donors in the form of ATP or GTP are required at several steps along the way. Gil et al. (2004) include 15 genes dedicated to nucleotide biosynthesis, while earlier work from Castellanos et al. (2007) lists only 12 genes. The discrepancy comes from what each author chose to include under the umbrella of 'nucleotide metabolism'. The study by Castellanos et al. (2004), while functionally complete in terms of the reactions necessary to synthesize specific nucleotides, neglected some aspects related to those reactions. For example, the action of ribonucleotide reductase (*nrdE* and *nrdF*) is coupled to the activities of thioredoxin (*trxA*) and thioredoxin reductase *trxB*, yet *trxA* and *trxB* were not included (Castellanos et al., 2004). Furthermore, Gil et al. (2004) included *prsA* for the synthesis of PRPP, whereas Castellanos et al. (2004) left that contribution coarse-grained. The study by Gil et al. (2004) took a holistic view of metabolism and thus the list of genes proposed by Gil et al. (2004) has been used here. More detailed diagrams of the nucleotide biosynthesis reactions are in Figures C.5 and C.6.

4.14.5 Cofactor Metabolism

A cofactor is a nonprotein chemical species that is required for an enzymes activity. These molecules, such as CoA, NAD^+, and FAD are essential for a functioning cellular metabolic network, but they are often overlooked or assumed constant in models of cellular metabolism. In designing the minimal

gene set, Gil et al. (2004) assumed that the cell has free availability of cofactor precursors, and that those precursors can enter the cell by simple diffusion. This assumption is followed for the precursors of cofactor metabolism. Once the precursor molecules are in the cytoplasm, cofactors can be synthesized via short biosynthetic pathways that minimize the cell's gene requirement. A representation of the cofactor biosynthesis reactions is in Figures C.7 and C.8.

4.14.6 Energy Metabolism and Fermentation

The minimal cell obtains ATP from substrate-level phosphorylation in glycolysis, producing lactate as the final step. In the process of producing lactate, NADH is reoxidized to NAD^+ and thus this bacteria is fermentative. Even small traces of oxygen are toxic to many anaerobic bacteria. When these strict anaerobes encounter oxygen, toxic products such as hydroxyl radical ($OH\cdot$) and hydrogen peroxide (H_2O_2) are formed. Hydrogen peroxide is dangerous to microorganisms because it can oxidize transition metals to form more hydroxyl radicals (White, 2000). Oxygen-tolerant microbes have the enzyme superoxide dismutase that catalyzes the transfer of the extra electrons from a radical oxygen to a second radical, forming hydrogen peroxide. These microbes also have the enzyme catalase, which breaks two hydrogen peroxide molecules into water. The proposed minimal gene set does not contain either of these enzymes (Gil et al., 2004). Therefore, it is assumed that the MCM represents a strictly anaerobic cell that exists in a benign, O_2-free environment.

The cell also produces energy in the form of a proton-motive force. As it is defined, the Gil et al. (2004) minimal gene set produces a proton-motive force by

exporting H^+ at the expense of ATP using the ATP synthase running in reverse. However, there is another possible source of proton export via the lactate exporter included in the MCM. While the actual proton export stoichiometry depends on the ΔpH, it has been estimated that this transporter exports one H^+ per lactate exported (Konings et al., 1994; Konings, 2002). This is not enough to generate a proton motive force to drive the uptake of inorganic nutrients, some amino acids, or to drive the ATP synthase in the forward direction to produce ATP. Thus, ATP synthase is included to maintain the protein gradient in the MCM.

In the Energy module of the MCM, the proton export rate associated with lactate transport is calculated. Because the cell drives a large amount glucose through glycolysis, lactate is produced and exported in large quantities (as may be expected with a fermentative bacteria). The protons required to drive nutrient transport are estimated, and the balance of the proton motive force is used to drive the ATP synthase and produce ATP. If the proton motive force is insufficient to drive nutrient uptake, the ATP synthase in the MCM will run in reverse and consume ATP.

4.14.7 Specific Reaction Notes

Cmk/Tmk

In the reaction scheme for the previously proposed minimal gene set shown in Figure 2 of Gil et al. (2004), CMP is not a necessary metabolite. However, it is produced in lipid synthesis by the PssA reaction (CTP is consumed to make phospholipid intermediates). The previous work is at some points contradictory

on the role that CMP will play. Gil et al. (2004) postulates that Tmk (EC 2.7.4.9) can perform the function of Cmk (EC 2.7.4.14), and therefore the *cmk* gene is not included. However, Figure 2 of (Gil et al., 2004) does reference the Cmk protein. Furthermore, in their follow-up work, Gabaldón et al. (2007) do include Cmk, claiming that it was omitted from Gil et al. (2004). Gabaldón et al. (2007) goes on to show that the NDK5 activity of Ndk is dispensable in their reaction network. However, this conclusion depends on the presence of a cytidylate kinase activity such as Cmk.

It was the original intention of Gil et al. (2004) was that *cmk* not be included in the minimal gene set based on the fact that several prokaryotes with reduced genomes use a single kinase to phosphorylate all pyrimidine nucleoside monophosphates (R. Gil, University of Valencia, personal communication, March 22, 2010). Therefore, Cmk is not included in the MCM, and the functions of the Cmk enzyme are fulfilled by Tmk.

CTP synthase

CTP synthase (EC:6.3.4.2) (coded by the *pygG* gene) is the enzyme that catalyzes the conversion of UTP to CTP with the addition of an amino group. The amino donor can be either -NH_3 or glutamine (KEGG reactions R00571 and R00573, respectively). Because the MCM explicitly encodes glutamine but not -NH_3, we opt to use the later reaction.

4.14.8 Reaction Reversibility

Some enzyme catalyzed reactions are thermodynamically reversible. There are two ways for us to treat reaction reversibility in this modeling framework. Some weakly reversible reactions can be approximated using inhibition terms. The effective rate decrease caused by an inhibition term at high product concentration will mimic the effect of a reverse reaction. The second way is to explicitly introduce a reverse reaction that has the opposite stoichiometry and a manually determined rate. The rate constant calculation procedure described in Section will set reverse reaction rates to zero unless they are explicitly given a lower bound (see Section 4.7.3). Therefore, all the rate constants for reverse reactions are manually curated in this system.

In the model presented here, it is assumed that most metabolic reactions are irreversible. It is of note that imposing the condition of reversibility on the reactions in the MCM has not been necessary to make the model simulation work. In nature, a large motivation for reversing reactions is to obtain building blocks for metabolic pathways when certain precursors are not available. A minimal cell, however, will have all of its nutrients supplied in its optimally supportive culture environment. Therefore, the need for having reversible enzymatic reactions is greatly reduced. Reversing a reaction could even be deadly for a minimal cell because the product of that reaction may not be produced by other reactions. However, in some cases it may be possible to further reduce the size of the minimal gene set by allowing reactions to be reversible, and this is suggested as a route for further study (Gabaldón et al., 2007).

4.15 DNA Replication

A computer chromosome is automatically constructed from the 241 genes in the MCM's minimal gene set. Because evidence has suggested that gene order is not conserved across distantly related bacterial species, the genes in this model are ordered arbitrarily (Mushegian and Koonin, 1996a; Tamames, 2001). It is proposed to implement a more rigorous scheme for gene ordering in future work (Section 6.2). The initiation (Section 4.15.1) and termination (Section 4.15.3) of DNA replication are now discussed, as well as the DNA synthesis reaction in the MCM (Section 4.15.2).

4.15.1 Initiation of DNA Replication

We have previously proposed a model for DNA replication in *E. coli* that relies on the titration of ATP-activated DnaA protein molecules binding to the origin of replication (Atlas et al., 2008). The specific mechanism of initiation, however, varies amongst bacterial species (Konieczny, 2003; Kogoma, 1997), and there are multiple options open to a minimal cell (Gil et al., 2004). DnaA, the proposed initiation protein used in the model of *E. coli*, is absent in the proposed minimal genome, and the authors suggest that DNA replication can take place without an initiation protein under some conditions (Gil et al., 2004). One condition that may lead to this phenomenon may be the possession of a small genome.

Gil et al. (2004) propose that the recruitment and loading of a helicase at the DNA origin of replication requires a histone-like protein (HupA) to destabilize the nearby DNA duplex. Once the duplex is destabilized, the helicase DnaB

is able to attract the primase DnaG to the replication fork. After a threshold level of primase is present at the fork, DNA replication initiates, HupA, DnaB, and DnaG are released, and the replisome takes over the synthesis of the DNA molecule. The proposed gene set also includes a DNA gyrase to assist in unwinding over wound regions of DNA during replication. To simulate the effect of HupA, DnaB, and DnaG proteins sequentially binding the *OriC* until they reach a threshold value, events (see Section 4.9) that monitor when the protein increases above or decreases below its threshold value are introduced. The sequence of events being modeled is depicted in Figure 4.3. The threshold value for activation by HupA is set to 30 molecules, which is based on similar values used for a model of initiation by DnaA protein in the *E. coli* model (Atlas et al., 2008). Specific data were not available for the number of molecules of DnaB or DnaG that are necessary to initiate chromosome synthesis, so it is assumed that one helicase (DnaB) is necessary at each replication fork (for a total of two per initiation), and that four DNA primases (DnaG) are necessary to allow the polymerase to commence DNA strand synthesis. It is assumed that these molecules are prevented from binding to the *Ori* after initiation commences because the chromosome is unwound.

To model binding reaction rates for small numbers of molecules usually requires a stochastic approach. For the 30 molecules of DnaA in previous studies it was found that a deterministic approach was adequate (Browning et al., 2004), so that approach is used here for HupA. For DnaB and DnaG, there is probably error inherent in modeling the binding of so few molecules deterministically, but it is assumed that the binding of HupA is the rate-controlling step in the process, and therefore the DnaB and DnaG binding reactions' primary purposes are to capture the necessity of these products for

cell division. The binding reactions are modeled using simple mass-action kinetics.

In *E. coli* DNA must be methylated for initiation to occur. It has been proposed that methylation functions to prevent the cell from reinitiating DNA synthesis before the previous round has progressed sufficiently. The semi-methylated DNA is recruited to the cell-membrane, which prevents subsequent initiations. This effect was captured in our previous model (Atlas et al., 2008). Gil et al. (2004) do not include a DNA methylase in the DNA replication section of their gene set, but they do include a "poorly characterized" methyltransferase, *mraW*. We assume here that *mraW* codes for a DNA methyltransferase that remethylates DNA after a round of DNA replication completes.

The assumptions used for replication initiation are sufficient to control DNA replication in the MCM. Previously, a much more complex model of the control of DNA replication initiation in *E. coli* using the DnaA protein was published (Browning et al., 2004; Atlas et al., 2008). The model presented here has similar constraints, but is simpler. The more sophisticated mechanism in *E. coli* and other bacteria may exist because they have to respond to a more complex environment. To the extent justified by experimental or theoretical evidence, it would be possible to include the more complex model of DNA replication initiation presented in Atlas et al. (2008). The framework presented here would allow the adoption of a more complex model in future work.

Note that the model for DNA replication initiation used in the MCM does not attempt to simulate the physical structure of the genome, but the actual physical structure may be important (Echtenkamp et al., 2009). The initiation

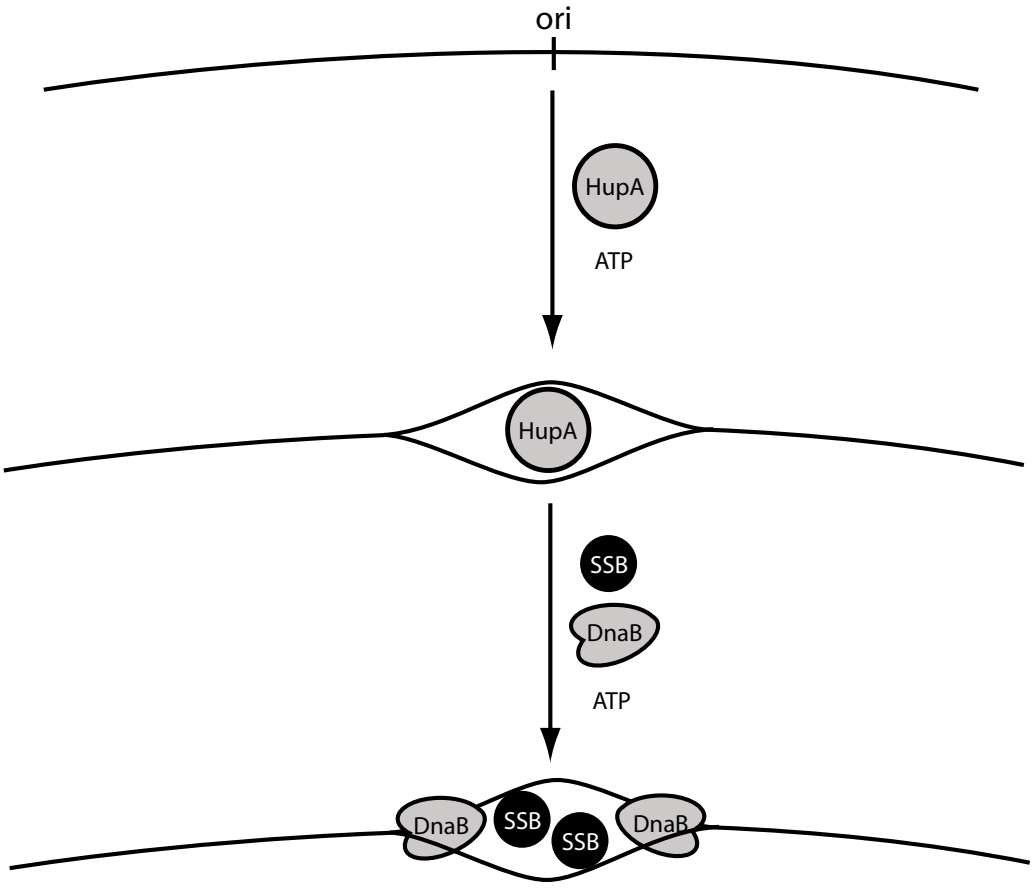


Figure 4.3: Mechanism for DNA replication initiation in the Minimal Cell Model. HupA is a histone-like protein, SSAB is Single-Stranded Binding Protein, DnaB is a helicase, and DnaG is a primase. In the proposed model, HupA destabilizes the DNA duplex near *Ori*, which allows the DnaB helicase and the DnaG primase into the replication fork. When DNA replication initiates, the proteins are released.

model is summarized in Figure 4.3.

The Gil et al. (2004) gene set does not include proteins for compaction or stabilization of the DNA structure, and it is assumed that these will be dispensable for bacteria with minimized genomes. This may be a weak assumption, but incorporating these genes would not alter the function of the mathematical model presented here as no description of the physical structure of the chromosome has been included.

4.15.2 DNA Synthesis

DNA synthesis is a coarse-grained reaction that consumes dNTPs in the relative proportions at which they are present in the chromosome. The reaction is catalyzed by the lumped "Replisome" protein, which contains protein products coded for by the $dnaE$, $dnaN$, $dnaQ$, $dnaX$, $holA$, $holB$, $gyrA$, $gyrB$, lig, and ssb genes. As the chromosome replicates, the relative distance along the chromosome (i.e. Fork Position) is calculated by comparing the amount of DNA synthesized to the mass of a single chromosome.

The mass of the chromosome is calculated from the sequences of all the genes in the minimal genome. In the current model the mass of the chromosome is $M_{CHR} \sim 3.77 \times 10^{-4}$ pg. The fork position is, therefore, calculated as,

$$Fork\ Position = \frac{M_3}{M_{CHR} \cdot num_{chrome}} - 1 \tag{4.37}$$

where M_3 is the mass of DNA in the cell, and num_{chrome} is the number of complete chromosomes. For these simulations, num_{chrome} is either one or two,

but it would be possible to allow multiple rounds of DNA replication initiation as in previous models developed for *E. coli* (Domach et al., 1984; Browning et al., 2004; Atlas et al., 2008).

4.15.3 Termination of DNA Replication

It is assumed that DNA replication terminates automatically when the replication fork reaches the DNA terminus, which consists of multiple copies of the TerA sequence from *E. coli* (Hill, 1992). After termination, the FtsZ protein is recruited to the midcell region to commence septum formation and division processes (Section 4.19.2). This process is perhaps the least mechanistic of these events and deserves attention in subsequent models.

4.16 Transcription

Individual genes are transcribed from the genome constantly throughout the cell cycle. Each RNA-coding locus on the chromosome has an RNA synthesis rate of the form in Equation 4.38.

$$\left(\frac{dRNA_i}{dt}\right)_S = v_{RNAi} \cdot \frac{GD_i}{GD_{sum}} \cdot \left(\frac{dM_2}{dt}\right)_S \tag{4.38}$$

$$\left(\frac{dM_2}{dt}\right)_S = \mu_{M2S} \cdot P2min_{sat} \cdot M_3 \cdot RNA_{pol} \tag{4.39}$$

In Equation 4.38 v_{RNAi} is a synthesis rate specific to RNA_i that is biologically

related to a promoter strength ($\frac{\text{pg RNA}_\text{i}}{\text{pg M}_2}$), $\frac{GD_i}{GD_{sum}}$ is the fraction of total gene dosage represented by gene i, and $\left(\frac{dM_2}{dt}\right)_S$ is the overall RNA synthesis rate for the cell (Equation 4.39). The gene dosage term appears for all mRNA synthesis equations by default, but if it is not required it can be optionally removed (i.e. when a gene's transcription is not regulated this way). In Equation 4.39, μ_{M2S} is the overall RNA synthesis rate constant ($\frac{\text{pg M}_2}{\text{h}\cdot\text{pg M}_3\cdot\text{pg RNA}_\text{pol}}$), $P2min_{sat}$ is a dimensionless saturation term for the scarcest ribonucleotide precursor, M_3 is the mass of DNA (pg), and RNA_{pol} is the lumped mass of enzymes involved in RNA synthesis (pg).

Note that due to the promoter strength constant in Equation 4.38, the sum of all RNA synthesis rates will not sum to $\left(\frac{dM_2}{dt}\right)_S$. Equation 4.39 is therefore supposed to represent a base capacity for RNA synthesis, the apportionment of which is determined for each RNA species by Equation 4.38.

Gene dosage for each gene is monitored automatically as a function of the replication fork position on the chromosome. If there is a single, non-replicating chromosome, in the cell, then the dosage for each gene is equal to its copy number. Once DNA replication begins, the gene dosage for each gene becomes a calculable function of fork position (fork position is defined in Equation 4.37).

There are two ways to calculate gene dosage. It can be updated via events each time the replication fork passes through a coding locus. For many genes, this tends to be a slow method because many events will fire as soon as the chromosome begins replicating. Alternatively, gene dosage can be calculated using a smooth function that approximates a step function. We use an exponential of the form shown in Equation 4.40.

$$HF(FP, gp) = \frac{1}{\left(1 + e^{-200 \cdot (FP - gp)}\right)} \tag{4.40}$$

where the heavy-step function (HF) is approximated as a function of the fork position (FP) and the position of a particular gene (gp).

It is important to verify that the synthesis rate of RNA and the approximate number of RNA polymerase molecules per model cell fall within reasonable ranges for natural bacteria. The combined molecular mass of the *rpoA*, *rpoB*, and *rpoC* gene products (and therefore of the RNA polymerase core enzyme) is 6.5×10^{-7} pg. These genes are included within a gene cluster, and it is estimated that the resulting RNA polymerase proteins account for about 50% of the protein products of this gene cluster, which corresponds to an average of about 5×10^{-3} pg of RNA polymerase per minimal cell, or approximately 6,550 molecules of RNA polymerase per cell, which falls in the range for *E. coli* (Bremer, 1996).

The transcription rate per molecule of RNA polymerase for stable RNA in *E. coli* is 85 $\frac{\text{nt}}{\text{s}}$ (Bremer, 1996), and the rate for mRNA in *E. coli* has been reported as 28-89 $\frac{\text{nt}}{\text{s}}$. If all the RNA polymerase molecules in the cell were active simultaneously, they could synthesize 0.3-1.1 $\frac{\text{pg}}{\text{h}}$ RNA, which is sufficiently above the 0.27 $\frac{\text{pg}}{\text{h}}$ RNA produced at the model cell's default conditions. Given the level of RNA polymerase in the MCM and the availability of precursors, the capacity of the cell to generate this level of RNA is sufficient.

It is assumed that RNA degradation is proportional to the mass of each RNA species in the cell. The rate constant for degradation can be set to a lower bound. Otherwise, it will be set to zero by the rate constant calculation procedure

because the procedure tries to minimize the sum of all reaction rate constants. Real cells require RNA degradation so they can reuse nutrients over the course of the cell cycle as different gene functions become necessary. For a minimal cell cultured under constant benign environment, the need for RNA turnover is far less compelling than for a cell that has a plethora of genes to choose from. Therefore, the MCM has relatively low degradation rate constants. Finally, it is assumed that "stable" RNA species such as ribosomal RNA (rRNA) have no degradation reactions.

4.17 Translation

Translation is governed by the following steps:

1. Production and maturation of rRNA species.
2. Production of ribosomal protein species.
3. Ribosome synthesis from ribosomal protein and rRNA species.
4. Production of 20 tRNA species.
5. Binding of the 20 amino acids to their corresponding tRNAs.
6. Protein synthesis with a stoichiometry based on the DNA/RNA sequence.

The overall process is depicted in Figure 4.4 and described in Sections 4.17.1 - 4.17.4.

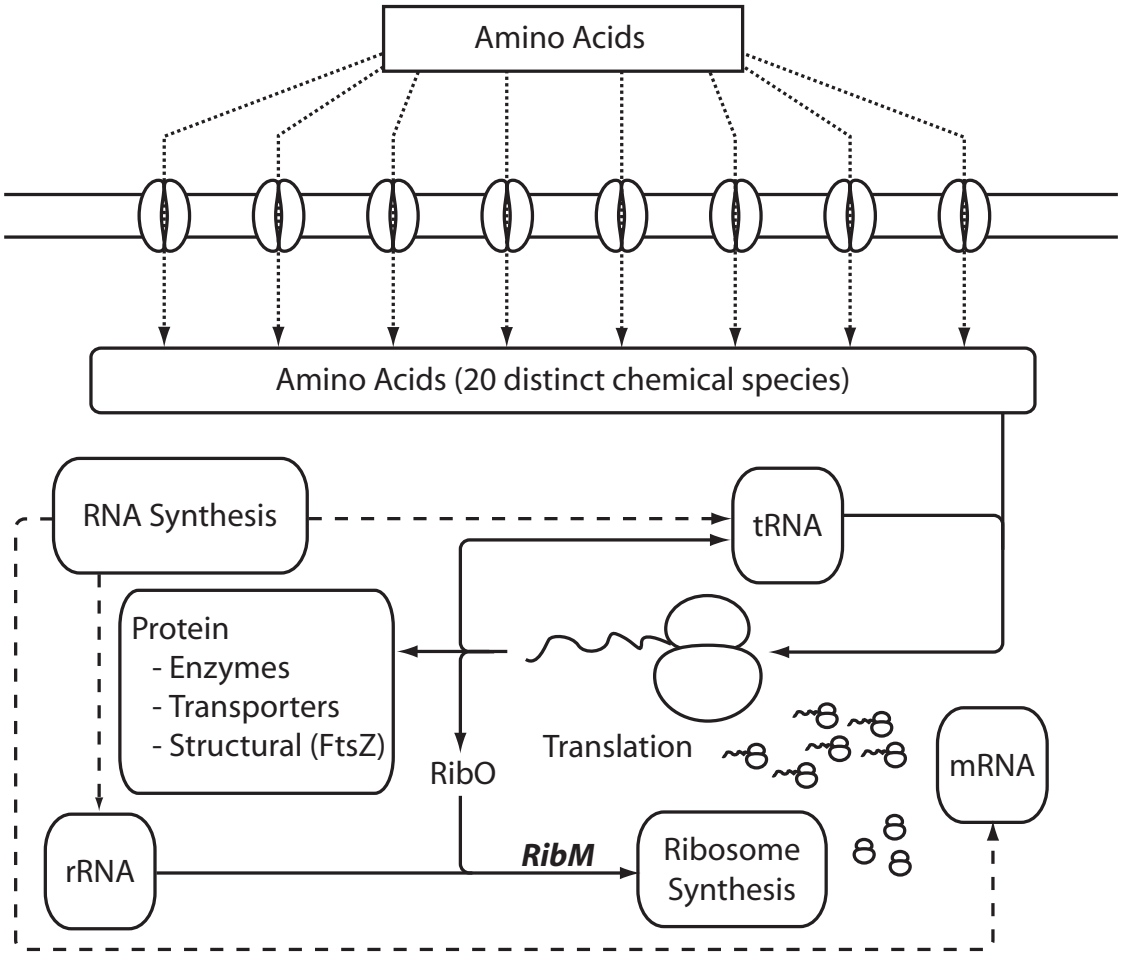


Figure 4.4: Protein synthesis scheme for the Minimal Cell Model. Solid arrows represent mass flow, while dashed arrows represent connections to other metabolic pathways or transport processes. Labels in *italic* are enzymes. Amino acids are imported into the cell through one of eight amino acid transport systems (see Table 4.7). The amino acids are combined with the appropriate tRNAs to form aa-tRNA species which proceed to ribosomes for protein synthesis. Note that tRNAs are recycled, and that some portion of protein synthesis (called RibO here) goes toward synthesizing ribosomal proteins.

4.17.1 Ribosome Synthesis

Prokaryotic ribosomes have a 50S (large) subunit and a 30S (small) subunit. Prokaryotes generally contain three rRNA molecules which are incorporated into ribosomes. The 50S subunit contains 23S and 5S rRNAs, while the 30S subunit contains a 16S rRNA. A wide range of prokaryotes have similar rRNA nucleotide compositions (Pace, 1973). In *M. genitalium* the 23S, 16S, and 5S rRNAs are coded for by the *rrlA*, *rrsA*, and *rrfA* genes, and their sequences are used in the model. It is of note that even though rRNAs are definitely required for cell growth, they are not included in the Gil et al. (2004) gene set because that list only includes protein-coding genes.

The rRNA species in this model are implemented under a single gene cluster that produces a species called $\mathrm{rti}_{\mathrm{RNA}}$, or "immature" rRNA. As in the previously published *E. coli* model (Domach and Shuler, 1984), the rRNA must mature before it is incorporated into ribosomes. The ribosome synthesis reaction combines mature rRNA with ribosomal proteins in the appropriate stoichiometry to form ribosomes.

4.17.2 Transfer RNA

Cells can have up to 61 unique codons in their genome. Each codon may pair with a different tRNA molecule, but many cells have fewer than 61 tRNAs. The Gil et al. (2004) minimal gene set does not include tRNA genes, however, tRNA is clearly a required part of protein synthesis and thus must be included in a minimal cell. It has been proposed that a minimal cell could survive with 21 tRNA species (one for each amino acid, and one for a start codon) if

the degeneracy were removed from the genomic code. In other words, if all degenerate codon expressions were collapsed to single codons, then the cell would only need 21 tRNAs.

On the other hand, the degeneracy of codon usage protects the cell in terms of robustness to DNA replication errors. If the codon usage were limited to 21, then the cell would be more prone to replication errors, and DNA repair mechanisms would become much more important.

For simplicity, we include 20 tRNA species in the MCM and assume that each species includes within it the tRNAs for all its corresponding codons. The model could be refined later to include more tRNA species. tRNAs combine with their corresponding amino acids through reactions to form amino acid - tRNA species. These reactions are catalyzed by the enzymes in the gene cluster mat_{tRNA}, which includes the *mnmA*, *mnmE*, *mnmG*, *rnpA*, *pth*, and *iscS* genes as proposed by (Gil et al., 2004).

4.17.3 Protein Synthesis

A translation model based on that used in the *E. coli* model is implemented in the MCM (Domach et al., 1984). Each protein's synthesis is of the form in Equation 4.41.

$$\frac{dM_{1i}}{dt}_{S} = \cdot k_i \cdot C_p \cdot frac_{rt} \cdot Rib_T \cdot ATP_{sat} \cdot M1p_{min\text{-}sat} \cdot \frac{mRNA_i}{M_{2M}} \cdot Trans_F \quad (4.41)$$

In Equation 4.41, M_{1i} is the mass of protein i (pg), k_i is the rate constant for the synthesis of protein i ($\frac{\text{pgM1}_i}{\text{pg } M1p\cdot\text{h}}$), C_p is the rate of protein

elongation ($\frac{\text{pgaminoacid}}{\text{h}}$) (Domach et al., 1984), $frac_{rt}$ is the fraction of actively translating ribosomes, Rib_T is the number of ribosomes in the cell, ATP_{sat} is a dimensionless ATP-dependent saturation term, $M1p_{min\text{-}sat}$ is a dimensionless saturation term based on the currently limiting amino-acyl tRNA, $mRNA_i$ is the mass of the mRNA for protein i (pg), $M2_M$ is the total mass of mRNA in the cell (pg), and $Trans_F$ is a lumped species representing the mass of all the non-ribosomal proteins involved in protein synthesis (pg). Note that the value of the mass of the limiting amino-acyl tRNA species, $M1p$ is defined Demand object (Section 4.18).

The stoichiometry of each protein synthesis reaction is determined by the DNA/protein sequence coded in the computer chromosome. The species consumed as reactants are actually the amino acid - tRNA species described in Section 4.17.2. In this manner, tRNAs are regenerated and reused as protein synthesis continues.

Methionine Aminopeptidase

The synthesis of proteins in all cells begins with methionine. During translation, the amino-terminal methionine of many proteins is cleaved by methionine aminopeptidase. For the majority of proteins in prokaryotes, the proteins that will have their methionine removed can be predicted by the amino acid in the second position in the sequence. The consensus is that if the penultimate amino acid is alanine, cysteine, glycine, proline, serine, threonine, or valine, then the leading methionine is likely to be cleaved (Sherman et al., 1985; Frottin et al., 2006).

The proposed minimal gene set does include a methionine aminopeptidase coded by the *map* gene (Gil et al., 2004). The activity of this gene has been included in the MCM by automatically adjusting the stoichiometry of protein synthesis for protein sequences that meet the criteria for cleavage. However, it should be noted that the MCM does not contain a detailed mechanism for protein synthesis. Thus, while the metabolic burden of synthesizing the Map protein is calculated, the concentration of the Map enzyme is not mathematically linked to protein synthesis.

4.17.4 Protein Degradation

Protein degradation is assumed to be proportional to the mass of protein in the cell. The rate constant for each protein's degradation rate was set to a lower bound of 0.025 h^{-1} in Domach. Because the protein degradation rate law is second order in the mass of protein and the mass of the protein degradation enzymes, we must choose a much higher rate constant with different units for any appreciable degradation to occur.

As a starting point, $1 \times 10^2 \left(\frac{\text{pg A degr}}{\text{pg A present} \cdot \text{pg Deg}_{\text{M1}} \cdot \text{h}}\right)$ is selected as a lower bound for the protein degradation rate constants, where A is the protein being degraded, and Deg_{M1} is the set of enzymes responsible for enzyme degradation.

4.18 Demands

Physiological processes such as DNA replication, transcription, and translation, consume many different reactants to create long biological polymers (i.e., DNA,

RNA, and protein, respectively). While it is possible to model a dependence on multiple substrates using a combination of Michaelis-Menten like saturation terms, the combination of many such terms leads to unreliable models because even if all the reactants are present in excess in the cytoplasm, the combination of many fractional terms can lead to greatly decreased reaction rates. For example, there are twenty reactants in the pseudo reaction that produces a particular protein product. Even at high concentrations, the cumulative effect of 20 saturation terms in a rate law could greatly decrease the calculated rate if they were all included.

Instead of including saturation terms for all reactants involved in these reactions, it is hypothesized that at any given time, a single reactant will have the highest "demand" in a reaction. We propose that synthesis of biological polymers only depends on single reactants in a Michaelis-Menten fashion. For example, translation will only depend on a single, limiting amino acid. During growth and development, particular amino acids will be more or less in demand and that single, limiting amino acid may not always be the same chemical species. To address that phenomenon, a 'Demand' class for was created the MCM. Each Demand object creates the parameters, equations, and events necessary to track the limiting reagent for a particular reaction. To create a Demand, one must specify the species that can act as limiting reagents for a reaction, as well as their saturation constant for that particular reaction.

The mass of each species is used to determine which chemical is in demand (i.e., the species with the lowest mass has the highest demand). This could later be updated to use the number of moles or molar concentration, but such an update is left as future work. The potential for demands to impact the

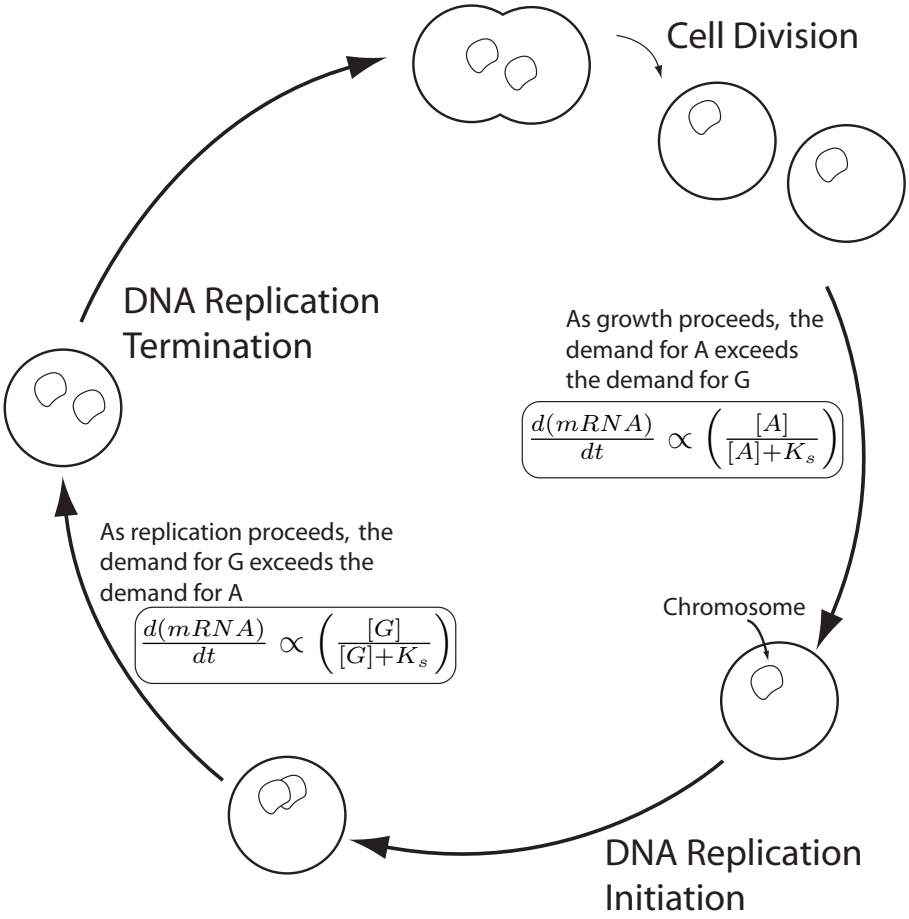


Figure 4.5: Chemical species demands over the course of the cell cycle. During the course of the cell cycle, changes in gene dosage can cause changing requirements for nucleotides. In this illustration, the demand is initially for ATP, and then switches to GTP.

cell behavior are illustrated in Figure 4.5, which shows an example of how the "in demand" species for a reaction could change over the cell cycle, and how that change affects the model equations. Note that at the beginning of the simulation, one (and only one) of the demand species in a Demand object must be limiting (i.e., the species associated with a particular Demand cannot all initially be equal). If they are, the system will not be able to select an initially limiting reagent.

The purpose of tracking this demand during the simulation is to calculate which reactant is limiting the reaction most severely at a given time. A high demand corresponds to a low concentration of a species, and a low demand corresponds to a high concentration. When the demand for species A surpasses the demand for species B, the reaction in question will automatically start using the mass of the more-limiting species in the calculation of the reaction rate.

4.19 Geometry

The shape of the model cell is determined automatically from the volume of its compartments (Sections 4.4.1,4.4.2). It is assumed that the cell shape is spherical (Figure 4.6). The two parameters describing the shape of the cell are the length of the cylindrical cell body (CL) and the width of the cell body (CW). For a spherical cell CL is always zero. The length of a dividing cell's dividing region (the septum) is referred to as SL. When termination of DNA replication completes and the cell division process starts, the enzyme FtsZ recruits membrane material to the septum. This results in a 'figure-eight' shaped cell where the connecting region gets thinner and thinner until the cell divides,

as in Figure 4.6(B). The current release of the MCM assumes a spherical shape by default.

4.19.1 Cell Volume

Given this shape, we can write the expressions for the surface area of the cell, (SA), in a cylindrical or spherical cell (Equations 4.42 and 4.43, respectively).

$$SA = \pi CW^2 + \pi CW \cdot CL \tag{4.42}$$

$$SA = \pi CW^2 \tag{4.43}$$

These expressions are true only before division has started (i.e. when no septum has formed). However, the surface area of the cell is also calculable from the mass of the cell membrane (Equation 4.44).

$$SA = f_S \cdot M_4 \tag{4.44}$$

where $f_S = 1.2 \times 10^2 \frac{\mu\text{m}^2}{pg}$ is the conversion factor for mass to surface area for the cell membrane based on *E. coli* (Domach and Shuler, 1984) and M_4 is the mass of the cell membrane (pg).

The mass of the cell membrane is used to calculate the SA so that Equation 4.42 can be rearranged to obtain an expression for the cell length in rod-shaped bacteria.

$$CL = \frac{SA - \pi CW^2}{\pi CW} \tag{4.45}$$

Equation 4.45 is still in terms of the cell width. To obtain cell width, we write an expression for the total volume of the cell depicted in Figure 4.7.

$$V = \frac{\pi}{6} CW^3 + \pi \frac{CW^2}{4} CL \tag{4.46}$$

Substituting the expression for CL in Equation 4.45 into Equation 4.46 yields:

$$V = \frac{\pi}{6} CW^3 + \pi \frac{CW^2}{4} \cdot \left(\frac{SA - \pi CW^2}{\pi CW} \right) \tag{4.47}$$

Equation 4.47 is an expression for the cell volume solely in terms of CW. We can solve for the width of the cell setting the expressions for volume in 4.3 and 4.47 to be equal. This modeling structure is an algebraic rule. The derivation for a spherical cell simpler because the cell does not have a body length (i.e., $CL = 0$), and Equation 4.46 reduces to Equation 4.48.

$$V = \frac{\pi}{6} CW^3 \tag{4.48}$$

4.19.2 Cell Division

The Gil et al. (2004) minimal gene set includes a single gene devoted to cell division, *ftsZ*. FtsZ is thought to be a major component of cytoskeletal structure, as well as a GTP binding protein and GTPase (Bramhill, 1997). The cell

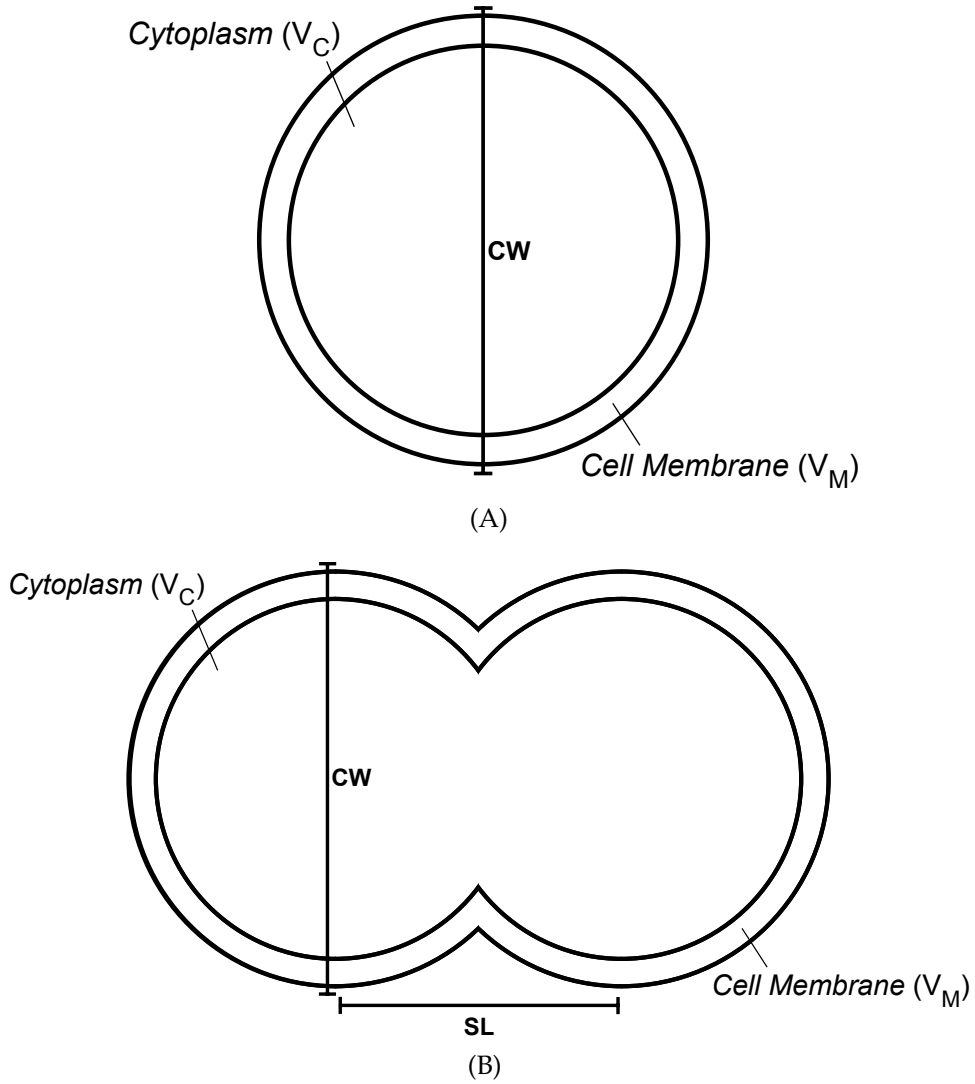


Figure 4.6: The spherical Minimal Cell Model. CW - Cell Width. The two labeled compartments, cytoplasm (V_C) and cell membrane (V_M), together comprise the volume of the whole cell, V. (a) The cell before septum formation begins. (b) The cell after septum formation as started. When the septum is complete (i.e. $SL = \frac{CW}{2}$), division occurs.

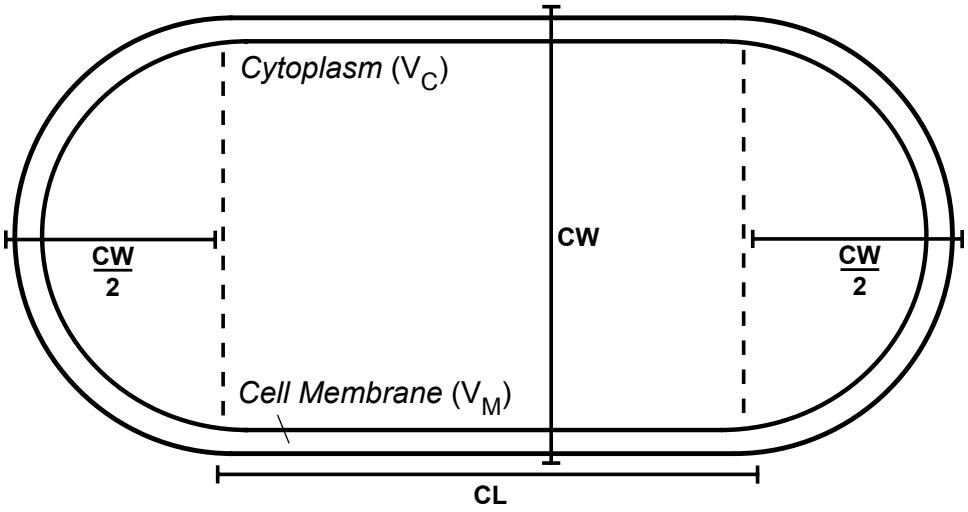


Figure 4.7: The cylindrical Minimal Cell Model. CW - Cell Width, CL - Cell Length. The two labeled compartments, cytoplasm and cell membrane, together comprise the volume of the whole cell, V.

has between 5,000 and 20,000 FtsZ molecules (Bramhill, 1997). After DNA replication is complete, FtsZ forms a ring at the midcell division site. FtsZ can operate in the absence of a cell wall (Bramhill, 1997), which again makes it a good candidate for an MCM. *ftsZ* is conserved over a wide range of species, and is the only *fts* gene present in *Mycoplasma*, although the Gil et al. (2004) minimal gene set also includes *ftsY*. Thus FtsZ is included in the MCM as the chief promoter of cell division. Some chaperonins are also implicated in division, but it is assumed here that the rate of division is controlled solely by the FtsZ protein.

Once DNA replication has completed (termination), the FtsZ protein in the cell is recruited in to the septal ring to catalyze the transfer of membrane material to the midcell region (Lutkenhaus and Addinall, 1997; Bramhill, 1997). After division, the FtsZ is released from the midcell (which has become the cell end cap) and reserved for subsequent divisions. Cell division occurs when the growing septum reaches the size of the diameter of the cell. In other words, after DNA termination, the septum is continually synthesized until it crosses the entire cell, effectively resulting in two physically separate daughter cells. An improved MCM could include mechanisms for positioning FtsZ and initiating cell division.

4.20 Minimal Gene Set

The MCM implements a whole-cell dynamic model of a single cell that contains the minimal gene set described by authors Gil et al. (2004). The authors break their minimal gene set into five major categories:

1. Information Storage and Processing
2. Protein Processing, Folding, and Secretion
3. Cellular Processes
4. Energetic and Intermediate Metabolism
5. Poorly Characterized

The differences between the Gil et al. (2004) gene set and what is included in this base MCM are reconciled in Sections 4.20.1 - 4.20.5. In particular, the minimal gene set proposed by Gil et al. (2004) only considers protein-coding genes (it does not include tRNA or rRNA species). Furthermore, the authors assumed that the cell could import amino acids and inorganic ions (e.g., K^+ and Mg^{2+}) from the environment through diffusion, but it is likely that transporters will be required. Finally, the authors suggest that the cell will synthesize ATP exclusively through substrate-level phosphorylation via lactate fermentation, but they provide no mechanism for synthesized lactate to exit the cell. Therefore, three rRNA genes, 20 genes tRNA genes, 14 genes coding for amino acid transport systems, four genes for transport of inorganic ions, and one gene corresponding to a lactate transporter has been added to the MCM. This yields a total of 241 genes (Appendix B). Figure 4.2 shows an overview of the metabolic features of the MCM, and each metabolic module is detailed in Appendix C. Table 4.8 shows a summary of how many genes fall into particular functional categories in the MCM. A full listing of the genes in the MCM is presented in Table B.3.

Table 4.8: Summary of genes used in the Minimal Cell Model, listed by category.

Category	No. Genes
Basic DNA replication machinery	14
Basic transcription machinery	8
Biosynthesis of Cofactors	12
Biosynthesis of nucleotides	15
Cell division	1
DNA repair, restriction, and modification	3
Glycolysis	10
Lipid metabolism	7
Pentose phosphate pathway	4
Protein folding	5
Protein post-translational modification	3
Protein translocation and secretion	5
Protein turnover	3
Proton motive force generation	9
Ribosomal RNA (rRNA)	3
Transfer RNA (tRNA)	20
Translation factors	12
Translation: amino-acyl-tRNA synthesis	21
Translation: ribosomal proteins	50
Translation: ribosome function, maturation, and modification	7
Translation: tRNA maturation and modification	6
Transport	23

4.20.1 Information Storage and Processing

DNA Metabolism

The DNA replication and repair systems are less complex in *Mycoplasma* species than in bacteria with larger genomes (Labarère, 1992), and it is expected that a minimal bacterium would have a simple DNA replication system. Gil et al. (2004) state that the four basic steps of DNA replication are: (i) Recognition of the origin of replication by protein components, (ii) Recruitment of initiator proteins to the origin to promote initiation of replication, (iii) DNA synthesis along two forks on the circular chromosome, and (iv) Replication termination and the separation of the daughter chromosomes.

The mechanism for DNA replication initiation varies widely in different bacteria. The MCM combines concepts proposed by Gil (Gil et al., 2004) and those used in a DNA replication model simulated in previous research on *E. coli* (Browning et al., 2004; Atlas et al., 2008). The mechanism used in the MCM is discussed in Sections 4.15.1 - 4.15.3.

Gil et al. (2004) include 13 genes in the minimal gene set for the purpose of DNA replication. Of those, three (*dnaB*, *dnaG*, and *hupA*) are modeled explicitly as initiators of DNA replication, while the remaining 10 are included in the *replisome* gene cluster.

Gil et al. (2004) include three genes in the minimal gene set for the purpose of DNA repair, restriction, and modification. It is debatable whether a minimal cell would require these functions. Because the MCM exists in a totally benign environment the extent of DNA damage would be minimized. However,

because DNA polymerase is error-prone, some DNA damage may occur even in a benign environment. Therefore, the three genes suggested in Gil et al. (2004) (*nth*, *polA*, *ung*) have been included. However, because the MCM does not include a mechanism for DNA damage, the protein products of these genes have no mathematical impact on the cell behavior. Currently, their only impact is via the energy burden the cell experiences in their synthesis. It is possible that this model might serve as the basis for a cell model where DNA damage is relevant and should be dealt with. In that case, the three genes included for DNA repair would have a mathematical function.

RNA Metabolism

Gil et al. (2004) list eight genes as being necessary for the basic transcription machinery. Of these, seven are included in an RNA polymerase gene cluster. The remaining gene, *nusA* is used in transcription/translation coupling, and is therefore included in the gene cluster for translation factors.

The MCM takes 19 of the 21 proposed amino-acyl-tRNA synthesis genes and includes them explicitly. The remaining two, *pheS* and *pheT*, are the α and β subunits of a single amino-acyl-tRNA synthetase, so they are included as a single gene cluster.

The six genes Gil et al. (2004) list for tRNA maturation and modification are included in the MCM as a single gene cluster.

There are 50 ribosomal proteins included in the Gil et al. (2004) gene set. All 50 of these are included in a single gene cluster called $ribO$, the largest gene cluster by far. In absence of a detailed mechanistic model for ribosome assembly

and function, these genes must remain in a single cluster with a single product corresponding to ribosomal protein.

The seven genes listed for ribosome function and maturation are included as a single gene cluster called $ribM$. The product of this gene cluster catalyzes the rRNA maturation and ribosome synthesis reactions in the MCM.

All 12 genes listed as translation factors in the Gil et al. (2004) gene set are, along with *nusA* included as a single "translation factor" gene cluster called $trans_F$.

There are two genes that participate in RNA degradation in the Gil et al. (2004) gene set, *pnp* and *rnc*. They are included as a single gene cluster called deg_{RNA}.

4.20.2 Protein Processing, Folding, and Secretion

The minimal gene set proposed by Gil et al. (2004) includes two genes related to post-translational modification. One of these, *pepA*, was omitted from the MCM gene set because it is unclear how its product, aminopeptidase A/I, would be used in the minimal cell. Gil et al. (2004) included *pepA* because it was present in all of the genomes they considered. However, it is nonessential in both *E. coli* and *B. subtilis* (Gil et al., 2004). The other gene dedicated to post-translational modification in the proposed minimal gene set is *map*, which codes for methionine aminopeptidase (Gil et al., 2004). The *map* activity has been included as described in Section 4.17.3.

Five genes for protein folding, *dnaJ*, *dnaK*, *groEL*, *groES*, and *grpE* are

included in the Gil et al. (2004) gene set. Because protein folding is required in all cells, we have included these genes in the MCM as a single gene cluster. However, the MCM does not contain a protein folding submodel, so the products of the protein folding gene cluster do not impact the model simulation.

Finally, the three "protein turnover" genes proposed by the Gil et al. (2004) gene set, *gcp*, *hflnB*, and *Ion* are included as a single gene cluster that catalyzes protein degradation.

4.20.3 Cellular Processes

Cell Division

Gil et al. (2004) propose that the only gene necessary for cell division in their minimal cell is *ftsZ*, and this gene is explicitly included in the MCM. At the time of DNA replication termination, FtsZ catalyzes the transfer of membrane material to the midcell region, promoting cell division.

Transport

Gil et al. (2004) include four genes related to transport of nutrients into the cell. *pitA*, an inorganic phosphate transporter, is included explicitly in the MCM. The three genes coding for the phosphotransferase system (PTS), *ptsG*, *ptsH*, and *ptsI* are included as a single gene cluster.

4.20.4 Energetic and Intermediate Metabolism

Metabolic processes straightforward to represent in the coarse-grained modeling framework, as these reactions are the main basis for the previous cell models (Domach et al., 1984).

All 10 genes listed by Gil et al. (2004) for glycolysis are included explicitly in the MCM.

The nine genes included as part of the ATP synthase machinery are included as a single gene cluster in the MCM. It is presumed that the ATP synthase runs in reverse to extrude protons and maintain the proton gradient. This is common behavior amongst lactic acid bacteria (Hutkins, 1993). However, if enough H^+ is exported by the lactate efflux at the end of the fermentation, it is possible that the ATP Synthase will run in the forward direction and generate ATP.

The four genes included for the pentose phosphate pathway are included explicitly in the MCM (Gil et al., 2004; Gabaldón et al., 2007).

The minimal gene set contains genes for synthesizing ATP through substrate level phosphorylation *only*. Specifically, the cell does not have an electron transport chain. It does contain the F1 ATPase in the cell membrane, but Gil et al. (2004) propose that this will run in reverse to help maintain a proton gradient.

The Gil et al. (2004) gene set does not explicitly address the issue of cellular use of NAD^+ vs. $NADP^+$ in terms of reducing power. A review of the reactions catalyzed by the minimal proteome reveals that in principle NAD^+ coupled with NADH should be sufficient. The single exception is that TrxB (thioredoxin reductase) does prefer $NADP^+$, but there is some evidence that a similar enzyme

could function with NAD^+ (Reynolds2002), so we follow the assumption of Gil et al. (2004) and Gabaldón et al. (2007) and use NAD^+/NADH for redox reactions. It is important to know whether the cell is capable of balancing redox species use. The metabolic rates in the MCM are able to balance NAD+ and NADH so that there is sufficient reducing power generated without an imbalance.

Of the seven genes listed for lipid metabolism, four (*cdsA*, *gpsA*, *psd*, and *pssA*) are included explicitly as single genes. The remaining three (*plsB*, *plsC*, and *fadD*) are included as a single gene cluster involved in lipid biosynthesis. *plsB* and *plsC* have been proposed as the basis for lipid membrane synthesis in semisynthetic minimal cells (Kuruma et al., 2009).

All 15 genes listed for nucleotide biosynthesis by Gil et al. (2004) are included explicitly as single genes in the MCM. The 12 genes list in Gil et al. (2004) for cofactor biosynthesis are also included in the MCM.

4.20.5 Additional Genes

The Gil et al. (2004) gene set proposes only four genes related to transport of nutrients into the cell, proposing that the cell should be able to obtain what it needs from the environment by diffusion (Gil et al., 2004). This may suffice for some nutrients, but it is likely that protein transporters will be necessary for many other nutrients. Therefore, the gene set proposed by Gil et al. (2004) is supplemented with an additional 19 genes dedicated to the transport of chemicals such as amino acids. The MCM has a total of 23 genes related to transport, which are listed in Table B.3.

The Gil et al. (2004) gene set neglects to mention coding regions for tRNA or rRNA species because they are not protein-coding genes. These genes, however, are clearly essential parts of the minimal genome for a modern chemoheterotrophic bacterium. The computer chromosome was supplemented with coding regions corresponding to 20 tRNA species. In cases where there are multiple tRNA alleles corresponding to a single amino acid, it is assumed that the tRNA region is actually a gene cluster coding for all of those alleles. The genome was also supplemented with genes for three rRNA species.

Large amounts of lactate are generated by the model because while the Gil et al. (2004) gene set includes lactate dehydrogenase, which consumes pyruvate and NADH, there is no reaction in the model that consumes lactate. We propose the addition of the *lctP* gene for export of lactate to the external environment.

4.20.6 Other Departures from the Proposed Minimal Gene Set

There are other genes that, while necessary for a minimal cell, have no mathematical model available for their interaction with the whole-cell. In these cases, we have elected to include the genes to ensure that we're accounting for their metabolic burden on the cell, but their genes and gene-products still have no connection to the rest of the cell. The mathematical model could be adjusted to show their function in future work. These genes include those whose gene products degrade macromolecules (deg_{M1} and deg_{RNA}), act solely on ions in the cell (kup, $mgtA$, $mntH$, $nhaB$, $pitA$, pmf, and ppa), or catalyze processes for which the MCM lacks any mechanistic details (dna_{rep}, $prot_{fold}$, map). The implications of these exceptions are discussed in Section 5.5.

The proposed minimal gene set includes the *pepA* aminopeptidase. However, there is no clear function for this gene in the minimal cell, so we choose not to include it. Eight "poorly" characterized genes are included in the gene set proposed by Gil et al. (2004) (see Table B.2). Most of these have no known function, but were included because they were present in all of the genomes considered in the study. Of these eight, only *mraW* is included in the MCM. MraW is a methyltransferase which is assumed to be necessary for DNA methylation and chromosome replication. However, the rest have no clear function for a minimal cell, and are therefore not included in the MCM.

The full list of genes from the gene set proposed by Gil et al. (2004) which have been excluded in the MCM is presented in Table B.2.

4.20.7 Analysis of the Minimal Gene Set

The minimal gene set proposed by Gil et al. (2004) has been analyzed in subsequent work by Gabaldón et al. (2007). To perform a structural analysis, Gabaldón et al. (2007) eliminated many of the 206 protein-coding genes from the minimal gene set proposed by Gil et al. (2004). Specifically, they removed polymerization reactions and any reactions involving macromolecules. Furthermore, they only considered reactions represented in the pathway maps of the KEGG database, which eliminates many reactions involving cofactors. They also only considered reactants and products what have at least one carbon atom in common on each side of the reaction. A metabolic reaction network was thus constructed by comparing the gene functions from Gil et al. (2004) to the new reaction database created in Gabaldón et al. (2007).

The connection degree distribution, clustering coefficient, average path length, and network diameter, were measured for the metabolic reaction network (Gabaldón et al., 2007). It was found that the average path length and network diameter tended to decrease with the size of the network (n) rather than with the size of the genome. An average path length and network diameter of 5.34 and 18, respectively, were reported for the minimal gene set (Gabaldón et al., 2007) when they considered a network with 165 nodes by applying the eliminations discussed above. Gabaldón et al. (2007) also found that a random network had a much smaller clustering coefficient than the natural or minimal gene sets ($C = 0.031$ for the minimal gene set compared to $C_r = 0.00977$ for a random network of the same size). However, the ratio C/C_r increases linearly with the number of nodes in a network, so smaller networks (including the minimal gene set) have less clustering. Most importantly, the results from Gabaldón et al. (2007) show that the minimal gene set and its corresponding reaction network behave as one would expect for a natural genome of the same size.

Gabaldón et al. (2007) also considered a reduced theoretical reaction network containing only 39 genes with 50 enzymatic steps for stoichiometric analysis. Their stoichiometric analysis did not include cofactor metabolism because, they argued, coenzymes play a catalytic function and do not affect the stoichiometric analysis. The reduced theoretical reaction network also assumes lactate to be a "sink" chemical whose concentration is essentially buffered.

Using the reduced theoretical reaction network, they investigated the robustness of the minimal gene set. They found that most mutations had a limited effect on the *topology* of the network, but that the removal of a few key

Table 4.9: Characteristics of the Minimal Cell Model genome.

Characteristic	MCM Value	Lit. Value	Reference
Genome size (kbp)	233	580	Value from *M. genitalium* (Fraser et al., 1995)
GC Content (%)	40	27.73	Median value for mollicutes (Sirand-Pugnet et al., 2007)
Gene density (%)	100	81-92	Various *Mycoplasma* species (Sirand-Pugnet et al., 2007)

enzymes had much more drastic effects. At the same time, the network was sensitive to sustained random attacks. This analysis, however, does not imply that the minimal gene set could be further reduced because maintaining the topology of a network is different than maintaining its viability (Gabaldón et al., 2007).

The minimal gene set used in the MCM is a modified and supplemented version of that presented by Gil et al. (2004). This genome's characteristics can be compared to those of some naturally occurring small-genome bacteria as in Table 4.9. The mollicutes, a category of bacteria that tend to have small size and small genome, do not have a common general organization to their genomes (Sirand-Pugnet et al., 2007), but some of their features could be used as organizational baselines for the MCM. For example, some mollicutes display bias in the GC skew near the chromosomal replication origin and DNA replication initiation loci. Table 4.9 lists a gene density of 100% for the MCM. This is because the MCM has no non-coding regions of DNA. If one or more non-coding regions are deemed necessary to bacterial survival, they can be added to the MCM as genetic loci.

The genes in the minimal bacterial gene set are not in all bacterial species, and even when they are the sequence for the gene is not always known. The genomic sequences for the proposed gene set were almost exclusively downloaded from the KEGG website (`http://www.genome.jp/kegg/`). For each gene in the minimal gene set, we searched the KEGG database gene bank for the following list of organisms, in order:

1. (*Mycoplasma genitalium*) - mge
2. (*Escherichia colu*) - eco

3. (*Bacillus subtilis*) - bsu
4. (*Wigglesworthia brevipalpis*) - wbr
5. (*Buchnera aphidicola*) - bap, bab, buc
6. (*Blochmannia floridanus*) - bfl
7. (*Synechococcus elongatus*) - syc
8. (*Mycoplasma gallisepticum*) - mga
9. (*Cytophaga hutchinsonii*) - chu
10. (*Bacillus pumilus*) - bpu
11. (*Rhodobacter sphaeroides*) - rsp

Table B.1 shows how many gene sequences were used from each organism.

4.21 Model Implementation and Availability

Prior work on the MCM was implemented in C++ only; this made it difficult to share the model code with other investigators. The updated model is available in the Systems Biology Markup Language (SBML) (Hucka et al., 2003). This representation is advantageous because anyone using SBML compatible tools should be able to access the model. The SBML version of the model contains 408 species, 570 reactions, and 36 events.

4.21.1 Simulation

The MCM is a differential algebraic equation system (DAE) with discontinuities due to discrete physiological events (e.g., cell division). The full set of equations and parameters in Systems Biology Markup Language format as well as instructions for download and simulation are available online at `http://minimalcell.bme.cornell.edu`. The DAE is integrated numerically using SloppyCell, a Python software package for simulation and analysis of biomolecular networks (Gutenkunst et al., 2007a). SloppyCell has been applied to several biological systems of interest (Waterfall et al., 2006; Gutenkunst et al., 2007b,c). Significant updates have been made to SloppyCell as part of the current research to adapt it to simulating a model of this size and complexity. Specifically, an integrator that can treat systems with algebraic constraints was added to the program, and support was added for several previously unsupported features of the SBML specification.

SloppyCell automatically compiles the structures listed in Table 4.1 and creates a Reaction Network object which can be integrated to obtain time course data for any variable in the model. All model simulation results presented in this dissertation are generated by integrating the model from an initial condition until a stable cell-division *limit* cycle is reached. It is common to study how bacterial behavior changes at different steady-state growth rates, which is controlled by varying the external nutrient concentration. While we have done preliminary exploration of response to reduced glucose levels, only growth at saturating levels of glucose is necessary for a minimal cell.

4.21.2 Testing Framework

A model must meet certain requirements to be considered a valid model of a minimal organism. Several of these requirements are testable computationally. Using the Python Unittest framework (`http://docs.python.org/library/unittest.html`), a set of automated tests was implemented to verify that new versions of the model met all minimal cell requirements. Most importantly, we aimed to automatically verify that every version of the minimal cell model meets the following requirements:

1. Genetic Minimality - No gene should be included that the cell can live without. Every gene in a minimal cell is essential, by definition. Therefore, any gene that is removed should result in model failure. A series of tests were implemented that sequentially remove each gene in the model, and verified that the loss caused model failure. Exceptions to this criterion are discussed in Section 5.5.
2. Resource Minimality - No resource should be included that the cell can live without. While the minimal cell does live in an optimally supportive culture environment, it should not have nutrients in the medium that it can do without. These tests remove each nutrient in turn from the medium to ensure that its loss causes model failure.
3. Structure Tests - Another set of tests checks to make sure that rules, events, and other model structures are working as expected in the MCM. For example, for all time in the cell, the sum of all protein masses should equal the total mass of protein in the cell (M_1). Similarly, the total mass of the cell should equal the mass of the membrane plus the mass of the cytoplasm. Tests to verify the correct functioning of new model structures are also

implemented. For example, there is a test to ensure that a Demand object (Section 4.18) results in exactly one limiting reagent being assigned to each reaction at a given time.

The full suite of tests for the MCM will be described in detail at the supplemental website described in Appendix I.

4.22 Conclusions

We have shown for the first time that it is possible to test the hypotheses behind a minimal gene set using a chemically detailed, dynamic, whole-cell modeling approach. An MCM with 241 product-coding genes (those which produce protein or stable RNA products) is presented. This gene set expands on the minimal gene set proposed by Gil et al. (2004). We assert that this set is genomically complete and codes for all the functions that a minimal chemoheterotrophic bacterium would require for sustained growth and division.

The modeling structures used for designing the MCM have been presented. These compartments, chemical species, parameters, reactions, rules, events, constraints, functions, and genes are all, in part, based on similar structures present in SBML. Designing structures based on SBML allows us to easily export the model to an SBML file, making it portable to other researchers who may be interested in making use of the model.

Development of the MCM required implementing techniques for estimating initial conditions and reaction parameters. An estimation technique for

determining initial concentrations for all chemical species in the cell was developed. This required making significant assumptions about the starting concentrations of proteins and metabolites in the cell. To further refine the MCM, it would be useful to have more precise initial conditions, particularly for precursor metabolites, proteins, and mRNA species. Because the model has many hundreds of rate and saturation parameters, a procedure to determine all unspecified parameters has been developed. This method takes advantage of the fact that over the course of a steady-state cell-cycle, every chemical species in a cell must double its mass.

This is the first hybrid bacterial cell model that includes reactions that have many activating substrates (e.g., protein synthesis depends on the concentrations of up to 20 amino-acyl-tRNA species). To effectively treat the reaction rates for reactions with many substrates, we introduced the concept of "demand" objects, which automatically create model equations necessary to track the concentration of the most "in demand" substrate.

The Shuler group has expertise in bacterial cell models that include the effects of discrete physiological events. These events depend on the chemical and genomic detail presented by the model, resulting in a clear connection between genomic sequence and physiological processes including DNA replication, transcription, translation, and cell division.

The metabolic and transport reactions included in the MCM are introduced in Section 4.14. Detailed illustrations of these metabolic pathways are included in Appendix C. These reactions are all balanced with respect to total mass. At least to a reasonable approximation the MCM's metabolism is in balance with respect to redox potential and carbon flow.

The minimal gene set proposed by Gil et al. (2004) was described, and the modifications to create the biologically complete minimal gene set used in the MCM were also introduced. In particular, we describe the inclusion of 19 extra genes dedicated to transport of nutrients and waste, as well as 20 tRNA genes and three rRNA genes. We included all genes from the Gil et al. (2004) minimal gene set except six with poorly characterized function (*mesJ*, *ybeY*, *ycfF*, *yoaE*, *yqgF*, and *yraL*), and one whose purpose we determined to be unnecessary in a minimal cell (*pepA*). The concept of a gene cluster, or a set of genes whose products perform a closely related function, was introduced as a way to coarse-grain the treatment of gene products whose functions could not be distinguished at the resolution of the current model. This MCM is not unique in the sense that other minimal gene sets or parameter sets could be used for the simulation and still produce a viable cell.

The MCM functions indefinitely in a benign, steady-state environment. A cell faced with any challenge, such as nutrient depletion or the start-up of a cell culture from an inoculum, may require additional genes to achieve robust cell-division. The major significance of this work is that it shows, for the first time, that it is possible to build a chemically and genomically detailed model of a minimal bacterium using the principles of coarse-grained bacterial cell modeling and reasonable assumptions about the cell and its environment.

REFERENCES

Atlas, J. C., Nikolaev, E. V., Browning, S. T., and Shuler, M. L. (2008). Incorporating genome-wide DNA sequence information into a dynamic whole-cell model of *Escherichia coli*: application to DNA replication. *IET Syst Biol*, 2(5), 369–382. doi:10.1049/iet-syb:20070079.

Berkelaar, M., Eikland, K., and Notebaert, P. (2010). lpsolve - Open source (mixed-integer) linear programming system, version 5.1.0.0, `http://lpsolve.sourceforge.net/`.

Bramhill, D. (1997). Bacterial cell division. *Annual Review of Cell and Developmental Biology*, 13, 395–424. doi:10.1146/annurev.cellbio.13.1.395.

Bremer, D. P., H. (1996). *Modulation of Chemical Composition and other Parameters of the Cell by Growth Rate, in Escherichia coli and Salmonella: Cellular and Molecular Biology, F.C. Neidhart, Editor*. ASM Press.

Brown, K. S. and Sethna, J. P. (2003). Statistical mechanical approaches to models with many poorly known parameters. *Physical Review E*, 68(2).

Browning, S. T., Castellanos, M., and Shuler, M. L. (2004). Robust control of initiation of prokaryotic chromosome replication: essential considerations for a minimal cell. *Biotechnology and Bioengineering*, 88(5), 575–584. doi:10.1002/bit.20223.

Browning, S. T. and Shuler, M. L. (2001). Towards the development of a minimal cell model by generalization of a model of *Escherichia coli*: Use of dimensionless rate parameters. *Biotechnology and Bioengineering*, 76(3), 187–192.

Burguire, P., Auger, S., Hullo, M.-F., Danchin, A., and Martin-Verstraete, I. (2004). Three different systems participate in L-cystine uptake in *Bacillus subtilis*. *Journal of Bacteriology*, 186(15), 4875–4884. doi:10.1128/JB.186.15.4875-4884.2004.

Burkovski, A. and Krämer, R. (2002). Bacterial amino acid transport proteins: occurrence, functions, and significance for biotechnological applications. *Applied Microbiology and Biotechnology*, 58(3), 265–274. doi:10.1007/s00253-001-0869-4.

Castellanos, M., Kushiro, K., Lai, S. K., and Shuler, M. L. (2007). A genomically/chemically complete module for synthesis of lipid membrane in a minimal cell. *Biotechnology and Bioengineering*, 97(2), 397–409. doi:10.1002/bit.21251.

Castellanos, M., Wilson, D. B., and Shuler, M. L. (2004). A modular minimal cell model: purine and pyrimidine transport and metabolism. *Proceedings of the National Academy of Sciences of the United States of America*, 101(17), 6681–6686. doi:10.1073/pnas.0400962101.

Chai, Y., Kolter, R., and Losick, R. (2009). A widely conserved gene cluster required for lactate utilization in *Bacillus subtilis* and its involvement in biofilm formation. *Journal of Bacteriology*, 191(8), 2423–2430. doi:10.1128/JB.01464-08.

Courville, P., Chaloupka, R., Veyrier, F., and Cellier, M. F. M. (2004). Determination of transmembrane topology of the *Escherichia coli* natural resistance-associated macrophage protein (Nramp) ortholog. *Journal of Biological Chemistry*, 279(5), 3318–3326. doi:10.1074/jbc.M309913200.

Dandekar, T., Snel, B., Huynen, M., and Bork, P. (1998). Conservation of gene order: a fingerprint of proteins that physically interact. *Trends in Biochemical Sciences*, 23(9), 324–328.

der Werf, M. J. V., Guettler, M. V., Jain, M. K., and Zeikus, J. G. (1997). Environmental and physiological factors affecting the succinate product ratio during carbohydrate fermentation by *Actinobacillus* sp. 130Z. *Archives of Microbiology*, 167(6), 332–342.

Domach, M. M. (1983). *Refinement and Use of a Structured Model of a Single Cell of Escherichia coli for the Description of Ammonia-Limited Growth and Asynchonous Population Dynamics*. Ph.D. thesis, Cornell University.

Domach, M. M., Leung, S. K., Cahn, R. E., Cocks, G. G., and Shuler, M. L. (1984). Computer model for glucose-limited growth of a single cell of *Escherichia coli* B/r-A. *Biotechnology and Bioengineering*, 26(9), 1140. doi:10.1002/bit.260260925.

Domach, M. M. and Shuler, M. L. (1984). Testing of a potential mechanism for *Escherichia coli* temporal cycle imprecision with a structural model. *Journal of Theoretical Biology*, 106(4), 577–585.

Echtenkamp, P. L., Wilson, D. B., and Shuler, M. L. (2009). Cell cycle progression in *Escherichia coli* B/r affects transcription of certain genes: Implications for synthetic genome design. *Biotechnology and Bioengineering*, 102(3), 902–909. doi:10.1002/bit.22098.

El-Hag, A. H., Zheng, Z., Boggs, S. A., and Jayaram, S. H. (2006). Effect of pore size on the calculated pressure at biological cells pore wall. *IEEE Transactions on Nanobioscience*, 5(3), 157–163.

Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., et al. (1995). The minimal gene complement of *Mycoplasma genitalium*. *Science*, 270(5235), 397–403.

Frottin, F., Martinez, A., Peynot, P., Mitra, S., Holz, R. C., et al. (2006). The proteomics of N-terminal methionine cleavage. *Molecular and Cellular Proteomics*, 5(12), 2336–2349. doi:10.1074/mcp.M600225-MCP200.

Gabaldón, T., Peretó, J., Montero, F., Gil, R., Latorre, A., et al. (2007). Structural analyses of a hypothetical minimal metabolism. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 362(1486), 1751–1762. doi:10.1098/rstb.2007.2067.

Gil, R., Silva, F. J., Pereto, J., and Moya, A. (2004). Determination of the core of a minimal bacterial gene set. *Microbiology and Molecular Biology Reviews*, 68(3), 518–537.

Gutenkunst, R. N., Atlas, J. C., Casey, F. P., Kuczenski, R. S., Waterfall, J. J., et al. (2007a). SloppyCell, `http://sloppycell.sourceforge.net/`.

Gutenkunst, R. N., Casey, F. P., Waterfall, J. J., Myers, C. R., and Sethna, J. P. (2007b). Extracting falsifiable predictions from sloppy models. *Ann N Y Acad Sci*, 1115, 203–211. doi:10.1196/annals.1407.003.

Gutenkunst, R. N., Waterfall, J. J., Casey, F. P., Brown, K. S., Myers, C. R., et al. (2007c). Universally sloppy parameter sensitivities in systems biology models. *PLoS Comput Biol*, 3(10), 1871–1878. doi:10.1371/journal.pcbi.0030189.

Hill, T. M. (1992). Arrest of bacterial DNA replication. *Annual Review of Microbiology*, 46, 603–633. doi:10.1146/annurev.mi.46.100192.003131.

Hochstadt-Ozer, J. and Stadtman, E. R. (1971). The regulation of purine utilization in bacteria. III. the involvement of purine phosphoribosyltransferases in the uptake of adenine and other nucleic acid precursors by intact resting cells. *Journal of Biological Chemistry*, 246(17), 5312–5320.

Hucka, M., Finney, A., Sauro, H. M., Bolouri, H., Doyle, J. C., et al. (2003). The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4), 524–531.

Hucka, M., Hoops, S., Keating, S., Le Novre, N., Sahle, S., et al. (2008). Systems biology markup language (SBML) level 2: Structures and facilities for model definitions. *Nature Precedings*. doi:doi.org/10.1038/npre.2008.2715.1.

Hutkins, N. L., Robert W. Nannen (1993). pH homeostasis in lactic acid bacteria. *Journal of Dairy Science*, 76, 2354–2365.

Kanehisa, M. and Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1), 27–30.

Kehres, D. G. and Maguire, M. E. (2003). Emerging themes in manganese transport, biochemistry and pathogenesis in bacteria. *FEMS Microbiology Reviews*, 27(2-3), 263–290.

Kogoma, T. (1997). Stable DNA replication: interplay between DNA replication, homologous recombination, and transcription. *Microbiology and Molecular Biology Reviews*, 61(2), 212–238.

Konieczny, I. (2003). Strategies for helicase recruitment and loading in bacteria. *EMBO Reports*, 4(1), 37–41. doi:10.1038/sj.embor.embor703.

Konings, W. N. (2002). The cell membrane and the struggle for life of lactic acid bacteria. *Antonie Van Leeuwenhoek International Journal of General and Molecular Microbiology*, 82(1-4), 3–27.

Konings, W. N., Lolkema, J. S., and Poolman, B. (1995). The generation of metabolic energy by solute transport. *Archives of Microbiology*, 164, 235–242.

Konings, W. N., Poolman, B., and Vanveen, H. W. (1994). Solute transport and energy transduction in bacteria. *Antonie Van Leeuwenhoek International Journal of General and Molecular Microbiology*, 65(4), 369–380.

Korn, E. D. (1969). Cell membranes: structure and synthesis. *Annual Review of Biochemistry*, 38, 263–288. doi:10.1146/annurev.bi.38.070169.001403.

Kuruma, Y., Stano, P., Ueda, T., and Luisi, P. L. (2009). A synthetic biology approach to the construction of membrane proteins in semi-synthetic minimal cells. *Biochimica et Biophysica Acta*, 1788(2), 567–574. doi:10.1016/j.bbamem.2008.10.017.

Labarère, J. (1992). DNA replication and repair. In J. Maniloff, R. McElhaney, L. Finch, and J. Baseman, editors, *Mycoplasmas Molecular Biology and Pathogenesis*, pages 23–40. American Society for Microbiology, Washington, D.C.

Luisi, P. L., Oberholzer, T., and Lazcano, A. (2002). The notion of a DNA minimal cell: A general discourse and some guidelines for an experimental approach. *Helvetica Chimica Acta*, 85, 1759–1777.

Lutkenhaus, J. and Addinall, S. G. (1997). Bacterial cell division and the Z ring. *Annu Rev Biochem*, 66, 93–116. doi:10.1146/annurev.biochem.66.1.93.

Merlin, C., Gardiner, G., Durand, S., and Masters, M. (2002). The *Escherichia coli metD* locus encodes an ABC transporter which includes Abc (MetN), YaeE (Meti), and YaeC (MetQ). *Journal of Bacteriology*, 184(19), 5513–5517.

Miles, R. (1992). Cell nutrition and growth. In J. Maniloff, R. McElhaney, L. Finch, and J. Baseman, editors, *Mycoplasmas Molecular Biology and Pathogenesis*, pages 23–40. American Society for Microbiology, Washington, D.C.

Murphy, C. K. and Beckwith, J. (1996). Export of proteins to the cell enverlope in *Escherichia coli*. *Escherichia coli and Salmonella Cellular and Molecular Biology*, 1, 967–978.

Mushegian, A. R. and Koonin, E. V. (1996a). Gene order is not conserved in bacterial evolution. *Trends in Genetics*, 12(8), 289–290.

Mushegian, A. R. and Koonin, E. V. (1996b). A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 93(19), 10268–10273.

Neidhardt, H. E., Frederick C. Umbarger (1996). Chemical composition of *Escherichia coli*. *Escherichia coli and Salmonella Cellular and Molecular Biology*, 1, 13–16.

Nikolaev, E. V., Burgard, A. P., and Maranas, C. D. (2005). Elucidation and structural analysis of conserved pools for genome-scale metabolic reconstructions. *Biophysical Journal*, 88(1), 37–49. doi:10.1529/biophysj.104.043489.

Overbeek, R., Begley, T., Butler, R. M., Choudhuri, J. V., Chuang, H. Y., et al.

(2005). The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Research*, 33(17), 5691–5702.

Pace, N. R. (1973). Structure and synthesis of the ribosomal ribonucleic acid of prokaryotes. *Bacteriological Reviews*, 37(4), 562–603.

Palsson, B. O. (2006). *Systems biology : properties of reconstructed networks*. Cambridge University Press.

Powell, E. O. (1956). Growth rate and generation time of bacteria, with special reference to continuous culture. *Journal of General Microbiology*, 15(3), 492–511.

Quintero, M. J., Montesinos, M. L., Herrero, A., and Flores, E. (2001). Identification of genes encoding amino acid permeases by inactivation of selected ORFs from the synechocystis genomic sequence. *Genome Research*, 11(12), 2034–2040.

Ramos, H. C., Hoffmann, T., Marino, M., Nedjari, H., Presecan-Siedel, E., et al. (2000). Fermentative metabolism of *Bacillus subtilis*: physiology and regulation of gene expression. *Journal of Bacteriology*, 182(11), 3072–3080.

Razin, S. (1975). The *Mycoplasma* membrane. *Progress in Membrane Science*, 9, 257–312.

Riley, M., Abe, T., Arnaud, M. B., Berlyn, M. K. B., Blattner, F. R., et al. (2006). *Escherichia coli* K-12: a cooperatively developed annotation snapshot–2005. *Nucleic Acids Research*, 34(1), 1–9. doi:10.1093/nar/gkj405.

Rodwell, A. W. (1969). A defined medium for *Mycoplasma* strain Y. *Journal of General Microbiology*, 58, 39–&.

Sarsero, J. P., Wookey, P. J., Gollnick, P., Yanofsky, C., and Pittard, A. J. (1991). A new family of integral membrane proteins involved in transport of aromatic amino acids in *Escherichia coli*. *Journal of Bacteriology*, 173(10), 3231–3234.

Schmelter, T., Trigatti, B. L., Gerber, G. E., and Mangroo, D. (2004). Biochemical demonstration of the involvement of fatty acyl-CoA synthetase in fatty acid translocation across the plasma membrane. *Journal of Biological Chemistry*, 279(23), 24163–24170. doi:10.1074/jbc.M313632200.

Sherman, F., Stewart, J. W., and Tsunasawa, S. (1985). Methionine or not methionine at the beginning of a protein. *Bioessays*, 3(1), 27–31. doi:10.1002/bies.950030108.

Shu, J. and Shuler, M. L. (1991). Prediction of effects of amino-acid supplementation on growth of *Escherichia coli* B/r. *Biotechnology and Bioengineering*, 37(8), 708–715.

Shuler, M. L. (2005). Computer models of bacterial cells to integrate genomic detail with cell physiology. *Proceedings of the KBM International Symposium on Microorganisms and Human Well-Being, June 30-July 2005, Seoul Korea*.

Shuler, M. L. and Dick, C. (1979). A mathematical model for the growth of a single bacterial cell. *Annals of the New York Academy of the Sciences*, 326, 35–55.

Silver, S. (1996). Transport of inorganic cations. *Escherichia coli and Salmonella Cellular and Molecular Biology*, 1, 1091–1102.

Singer, S. J. and Nicolson, G. L. (1972). The fluid mosaic model of the structure of cell membranes. *Science*, 175(23), 720–731.

Sirand-Pugnet, P., Citti, C., Barr, A., and Blanchard, A. (2007). Evolution

of mollicutes: down a bumpy road with twists and turns. *Research in Microbiology*, 158(10), 754–766. doi:10.1016/j.resmic.2007.09.007.

Tamames, J. (2001). Evolution of gene order conservation in prokaryotes. *Genome Biology*, 2(6), research0020.10020.

Tamames, J., Gonzlez-Moreno, M., Mingorance, J., Valencia, A., and Vicente, M. (2001). Bringing gene order into bacterial shape. *Trends in Genetics*, 17(3), 124–126.

Waterfall, J. J., Casey, F. P., Gutenkunst, R. N., Brown, K. S., Myers, C. R., et al. (2006). Sloppy-model universality class and the vandermonde matrix. *Phys Rev Lett*, 97(15), 150601.

White, D. (2000). *The Physiology and Biochemistry of Prokaryotes*. Oxford University Press, New York, New York, 2nd edition.

Zheng, S. and Haselkorn, R. (1996). A glutamate / glutamine / aspartate / asparagine transport operon in *Rhodobacter capsulatus*. *Molecular Microbiology*, 20(5), 1001–1011.

CHAPTER 5

MINIMAL CELL MODEL APPLICATIONS

5.1 Introduction

This dissertation considers construction of a Minimal Cell Model (MCM) based on the gene set proposed by Gil et al. (2004). The MCM simulates a hypothetical bacterial cell with the minimum number of genes necessary to grow and divide in an optimal environment (Browning and Shuler, 2001). There are several applications of an MCM. One application is that it can act as a "learning model" used to test our understanding of biochemistry and microbiology; our ability to construct a chemically and genomically detailed model of a chemoheterotrophic bacterium tells us that our understanding of metabolism is not lacking anything essential. Additionally and practically, it can serve as a platform to test the effects of biochemical and genetic interventions on cell behavior. Furthermore, it has been proposed than an MCM is an important step toward the development of a synthetic platform cell for biotechnology (Foley and Shuler, 2010).

An MCM serves as a framework to test hypotheses about minimal bacterial cells as well as microbiology in general. The major contribution of the MCM presented in Chapter 4 is that it tests the plausibility of the proposed minimal gene set used to create it. It is shown for the first time that it is possible to create a genomically and chemically detailed model of a minimal cell that is capable of simulating sustained replication in an optimally supportive culture environment.

This chapter explores specific applications of the base model to probe its

general behavior and predictive capabilities. Section 5.2 demonstrates use of the MCM to calculate important bacterial growth parameters. The application of phase plane analysis to two pairs of model variables is shown in Section 5.3, and, in a related analysis, the movement of the position of a gene around the computer chromosome is considered in Section 5.4. In Section 5.5 we explore how manipulating the activity of a gene product or expression of a gene can affect the model cell's survival. Section 5.6 shows how an MCM could be used to aid development of nutrient media for small-genome synthetic cells, with a focus on the effects of competitive inhibition on transport systems with multiple substrates. Finally, in Section 5.7, the MCM's response to removing a particular activity of the Ndk protein is compared to previous results for a structural analysis of the minimal gene set proposed by Gil et al. (2004).

5.2 Calculation of Growth Parameters

Part of the utility of a chemically detailed cell model is that an engineer can design experiments that probe its behavior in response to various environmental and genetic manipulations. The MCM can also serve as a platform to evaluate and test the plausibility of candidate minimal gene sets, as it does in the work presented here. One way to perform such a test is to compare the model predictions to those for general chemoheterotrophic bacteria. While there is not an experimental lab-bench analog of the MCM, it is comparable to a generalized chemoheterotrophic bacterial cell (Browning and Shuler, 2001; Castellanos et al., 2004).

Table 5.1 shows calculated growth and molecular composition parameters

obtained using the MCM. These values are compared to values for *Escherichia coli* from Bremer (1996). In Table 5.1, genetic sequence measurements are based on values from *Mycoplasma* and other organisms listed in the KEGG database (Kanehisa and Goto, 2000). For a summary of organisms used as the basis of the MCM's gene set and the full list of genes, see Tables B.1 and B.3. Parameters in class I are inputs to the model (e.g., the number of deoxyribonucleotide residues per genome is fixed by the sequences of the genes in the minimal gene set). Parameters in classes II-V are outputs from the model simulation, except for C_p, which is an input constant based on a previous model of *E. coli* (Domach et al., 1984). The five classes in Table 5.1 are defined as:

I. Structural parameters that do not vary with growth rate. These parameters are calculated from the genome/proteome sequence of the minimal cell.

II. Partition factors which are essentially invariant. The values presented are typical values for the model and are close to those for *E. coli* presented by Bremer (1996).

III. Other partition parameters expected to vary with the growth rate. The values presented here are for a minimal cell with growth rate equal to 0.86 h^{-1}.

IV. Kinetic parameters describing functional activities. The peptide chain elongation rate, C_p, is a constant parameter of the model, which we chose to match the value used by Domach et al. (1984). The DNA chain elongation rate, C_d, is calculated by dividing the chromosome length by the length of time it takes to replicate the chromosome during the simulation (the C period).

V. Chromosome replication and cell division parameters calculated by the simulation.

There are many areas of agreement between the *E. coli* data and the MCM (e.g., fraction of active ribosomes, or DNA chain elongation rate). However, some calculations from the MCM do not match the data from *E. coli* due to the nature of a minimal cell. In class I, for example, the deoxyribonucleotide residues per genome will be lower in the MCM because it is a model of a cell *defined* by its low number of genes. Slight differences in the sequence lengths for ribosomes, tRNAs, and RNA polymerase occur due to sequence differences between *E. coli* and the source organisms used for the MCM. The partition factors (classes II and III) show strong agreement between *E. coli* and the MCM, and one would expect these features to hold constant amongst many bacterial species. The peptide chain elongation rate, C_p, is in agreement with the high-end of the values for *E. coli*, but this quantity is actually an input to the model based on data for *E. coli* (Domach, 1983), so it is unsurprising that they concur. The DNA chain elongation rate, which is calculated from the model simulation by dividing the chromosome length by the C-period length, falls significantly below that of *E. coli*. *Mycoplasma* species tend to have slow DNA replication rates, e.g. 100 bp/s in *M. capricolum* (Seto and Miyata, 1998), so it is not unexpected that a minimal cell would also have slower DNA replication rates. However, because of its minimized chromosome, the MCM actually exhibits a shorter C-period (24-25 minutes) than *E. coli*. Finally, the D-periods for the MCM and *E. coli* are similar (20.2 min for *E. coli* vs. 19.6 min for the MCM).

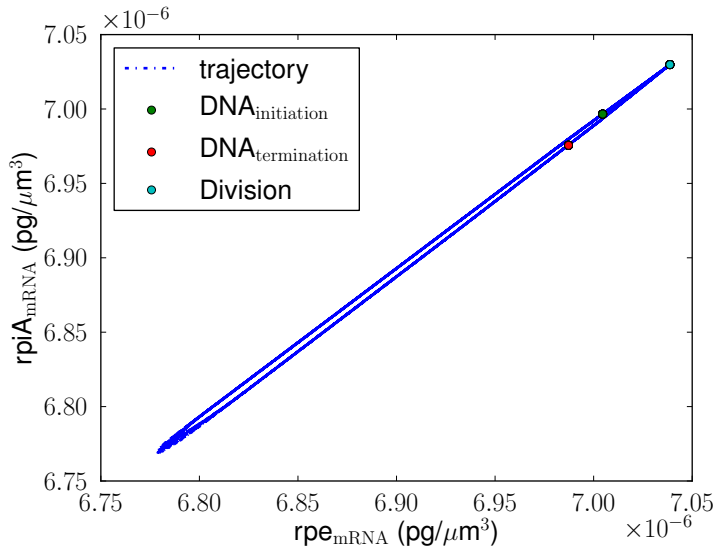
Table 5.1: Parameters related to the growth and molecular composition of the Minimal Cell Model. This table is modeled after Table 1 from (Bremer, 1996). See the main text for a definition of parameter classes I-V.

Class	Parameter	Symbol	*E. coli*	MCM
I	Deoxyribonucleotide residues per genome	kbp/genome	4700	233
	Ribonucleotide residues per 70S ribosome	nucl/prib	4566	4546
	Amino acid residues per 70S ribosome	aa/rib	7336	6856
	Ribonucleotide residues per tRNA	nucl/tRNA	80	77
	Amino acid residues per RNA polymerase core	aa/pol	3407	3010
II	Fraction of total RNA that is stable RNA	f_{sRNA}	0.98	0.96
	Fraction of stable RNA that is tRNA	f_{tRNA}	0.14	0.15
	Fraction of active ribosomes	$frac_{rt}$	0.921	0.797
III	Fraction of total protein that is r-protein	α_r	0.09-0.22	0.12
	Fraction of total protein that is RNA polymerase	α_p	0.009-0.01	0.03
IV	Peptide chain elongation rate	C_p	12-22 aa/s	23 aa/s
	DNA chain elongation rate	C_d	500-830 nucl bp/s	184 nucl bp/s
V	Time to replicate the chromosome	C	40-67 min	21.1 min
	Time between termination of replication and division	D	20.2 min	19.5 min

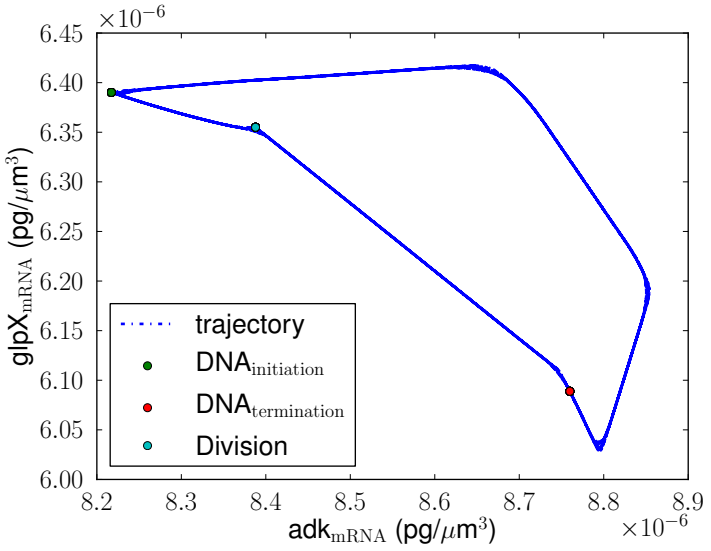
5.3 Phase Plane Analysis

A simulation of the MCM will produce time series mass data for every chemical species defined in the cell. This data can be plotted as time series trajectories that show the cell's approach to steady-state for each chemical species. One can also plot the concentrations of two different variables in the cell against each other. This technique is called 'phase plane' analysis. A well-studied example of phase plane analysis is the application to predator-prey systems (May, 1972). Considering dynamical systems two variables at a time is useful to visualize their behavior. In N-dimensional systems ($N > 2$), the technique projects the N-dimensional space onto a 2D-plane. Variables in the MCM can also be studied in the phase plane by allowing the system to approach a steady-state and then plotting two variables in the phase plane. Because the system is attracted to a closed trajectory, the curve on the phase plane is called a 'limit-cycle'.

Figure 5.1 shows phase plane plots for two pairs of mRNA species. The transcripts shown correspond to the genes *rpe* (ribulose-phosphate 3-epimerase), *rpiA* (ribose 5-phosphate isomerase), *adk* (adenylate kinase), and *glpX* (sedoheptulose-1,7-bisphosphatase). Genes that are not part of gene clusters were selected to limit the study to transcripts corresponding to individual genes. Recall that in the MCM chromosomal position is measured from 0.0 to 1.0, and that the chromosome position for each gene is arbitrary (but constant). One pair with adjacent chromosomal positions (0.531 and 0.535) and one pair with widely separated chromosomal positions (0.001 and 0.876) were selected. Figure 5.1(A) shows that for the pair with adjacent chromosomal positions (*rpe* and *rpiA*), the concentrations of each mRNA track with each other irregardless of cell cycle position. In contrast, the separated pair of genes



(A)



(B)

Figure 5.1: Phase plane analysis of mRNA species in the MCM. The compositions corresponding to DNA replication initiation, DNA replication termination, and cell division are shown on the plot. (A) shows a phase plane plot for mRNA products coded for by genes that have adjacent chromosomal positions (*rpe* and *rpiA*, located at 0.531 and 0.535, respectively). (B) shows a phase plane plot for mRNA products coded for by genes with widely separated chromosomal positions (*adk* and *glpX*, located at 0.001 and 0.876, respectively).

(*adk* and *glpX*) have transcript levels whose relative values change over the course of the cell cycle, as in Figure 5.1(B). The *adk* gene is located close to the origin of replication, and is copied soon after replication initiates. The plot shows a corresponding increase in adk_{mRNA} relative to $\text{glpX}_{\text{mRNA}}$. This continues until the *glpX* gene is copied, causing a change in the *glpX* gene dosage and eventually and increase in the $\text{glpX}_{\text{mRNA}}$ concentration. This implies that in the absence of other types of regulation, the chromosome position has a significant influence on transcript production, a result that is explored further in Section 5.4. This observation may be particularly relevant if the ratio of one gene product to another is physiologically important. It is possible to generate limit-cycle plots for any two variables in the MCM.

5.4 Gene Position Affects Protein Production

Another genetic manipulation addressed by the MCM is the effect of gene position on protein production. The MCM's computer chromosome is automatically constructed from the genes in its minimal gene set. There is conflicting evidence regarding the conservation of gene order in bacteria (Mushegian and Koonin, 1996; Dandekar et al., 1998; Tamames et al., 2001; Tamames, 2001). Because the prevailing evidence suggests that gene order is not conserved across long evolutionary distances in bacterial species (Mushegian and Koonin, 1996; Tamames, 2001), the genes are ordered arbitrarily in this first release of the full MCM.

Even though gene order is generally not conserved in bacteria, there is evidence in *E. coli* that transcript levels for a significant fraction of genes are

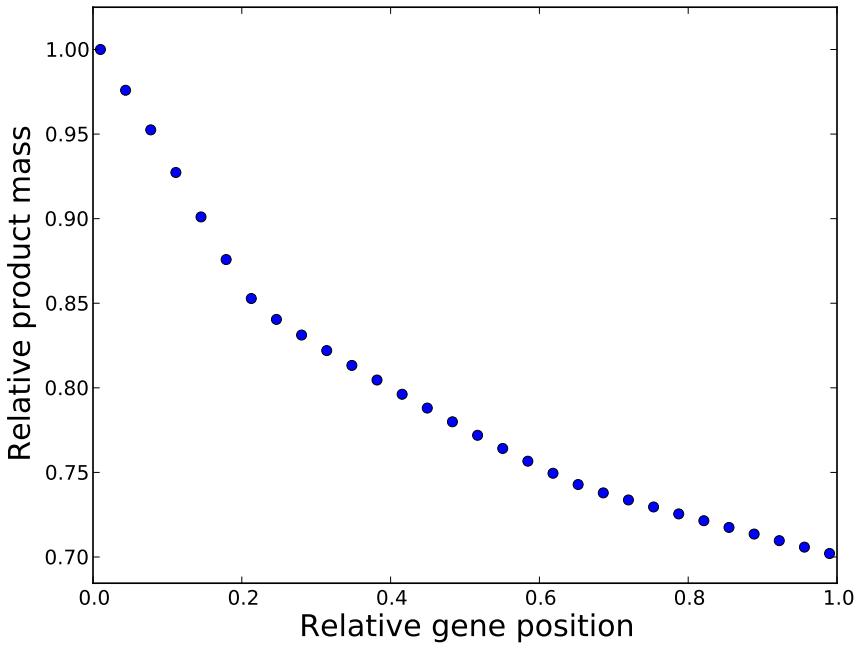


Figure 5.2: Effect of gene chromosomal position on protein product production for a hypothetical gene and its protein. Fork position represents the gene's position on a scale from 0.0 (the *Ori*) to 1.0 (the terminus of replication).

affected by cell cycle progression (Echtenkamp et al., 2009). To measure the impact of DNA replication progression in the MCM, a hypothetical gene *ins* (for *insert*) that codes for a hypothetical protein Ins was introduced to the computer cell. On successive simulations, the gene's fork position was varied from 0.0 to 1.0 (the cell's computer chromosome is normalized so that chromosome positions fall in this range). Figure 5.2 shows that the production of protein Ins dramatically decreases as the coding gene moves along the chromosome. This is the effect of gene dosage or gene copy number. As a gene is moved closer to the origin of replication, its average copy number over the course of the cell cycle is increased because it is replicated sooner. Those genes that

are farther along the chromosome are replicated later, and therefore have lower average copy numbers. The production of a particular protein is related to its corresponding mRNA levels, and the mRNA production depends on the gene dosage. In the absence of other types of regulation, proteins at the beginning of the chromosome have a higher production rate in the MCM. This result also has implications for synthetic cell design. If the position of a gene on the chromosome can affect its expression levels significantly, then chromosome design, including a rational choice of the relative positions of coding sequences, will have to be tightly coupled to cell design.

5.5 Knockout Experiments and Gene Essentiality

The MCM can be used to probe the effects of genetic or other manipulations on the cell's survival. The essentiality of each gene in the gene set was tested using knockout experiments. All gene and gene cluster knockouts in the model, except for 12, cause simulated cell death. Those that do not cause cell death correspond to genes whose products:

- degrade macromolecules (deg_{M1} and deg_{RNA}). A cell that is totally unchallenged may not actually require degradative pathways for macromolecules. One of the primary reasons for degradation is to recycle resources in a changing environment. The MCM is in an idealized constant environment, and does not depend on degradation for recycle of important precursors. That said, it is likely that in a real, even near-ideal situation, these enzymes would be necessary.
- act solely on ions in the cell (kup, $mgtA$, $mntH$, $nhaB$, $pitA$, pmf, and

ppa). These code for inorganic ion transporters, with the exception of ppa, which is an inorganic pyrophosphatase. The MCM does not track the concentrations of ions in the cell and assumes they are always available at sufficient concentration to satisfy cellular needs. These genes are included because (1) they are generally accepted as being necessary for cell survival, (2) it is important to track the energy and precursors consumed in the synthesis of their mRNA and protein products, and (3) the rates that their protein products operate at is calculated to provide an estimate of energy consumption related to transport processes. That said, removing these products does not result in cell death in MCM simulations because the ion concentrations that they affect are assumed to be buffered.

- catalyze processes for which the MCM lacks any mechanistic details. (dna_{rep}, $prot_{fold}$, map). dna_{rep} and $prot_{fold}$ are gene clusters that correspond to DNA repair, protein folding, respectively. The MCM does not contain a model for DNA damage, therefore it does not suffer from lack of a system for repairing damaged DNA. If a model for DNA damage were introduced, these gene products could be explicitly linked to the cell's survival. Protein folding is also a process that has no mechanistic detail in the MCM, but because protein folding is a feature of all cellular life, these genes were included to account for the metabolic burden of their expression. map corresponds to methionine aminopeptidase. This gene is included in the sense that methionine is cleaved from proteins when necessary, but the presence of the enzyme Map is not mathematically linked to the current MCM.

Any cell constructed on the lab bench would likely require all of the genes listed above because the assumption of a constant, benign environment can only

be approximated at best. In the context of the assumptions made here for a mathematical model these 12 genes are not essential.

It is possible that initial conditions could be an important factor in successfully synthesizing a minimal bacterial cell in the lab. We present here an example of how the initial conditions of the model can affect the output and robustness of the computer cell. First, the effect of enzyme mass on cell survivability is considered. Then, we consider how knockout interventions can change dynamics in the cell to the point where it dies.

Figure 5.3 considers the effects on ATP mass of reducing the activity of the phosphoglucose isomerase reaction (Equation 5.1), which is catalyzed by the Pgi enzyme according to the rate law in Equation 5.2.

$$\text{glucose-6P} \longrightarrow \text{fructose-6P} \tag{5.1}$$

$$\frac{dP}{dt}_S = v_{Pgi} \cdot \frac{g6P}{g6P + Ks_{g6P} \cdot V_C} \cdot Pgi \tag{5.2}$$

In Eqn. 5.2, P is the mass of fructose-6P, $g6P$ is the mass of glucose-6P, v_{Pgi} is the reaction rate constant ($\frac{mass\ P}{time \cdot mass\ E}$), Ks_{g6P} is a saturation constant describing the activating effect of glucose-6P on the reaction ($\frac{mass}{volume}$), V_C is the volume of the cytoplasm, and Pgi is the mass of enzyme Pgi. This reaction is the first step of glycolysis in the MCM, and is a bottleneck for energy metabolism and for producing the precursors of anabolic metabolism.

Figure 5.3-A shows the default trajectory for the mass (in pg) of ATP over time. The steep drop in ATP mass every ∼0.8 h corresponds to the moment

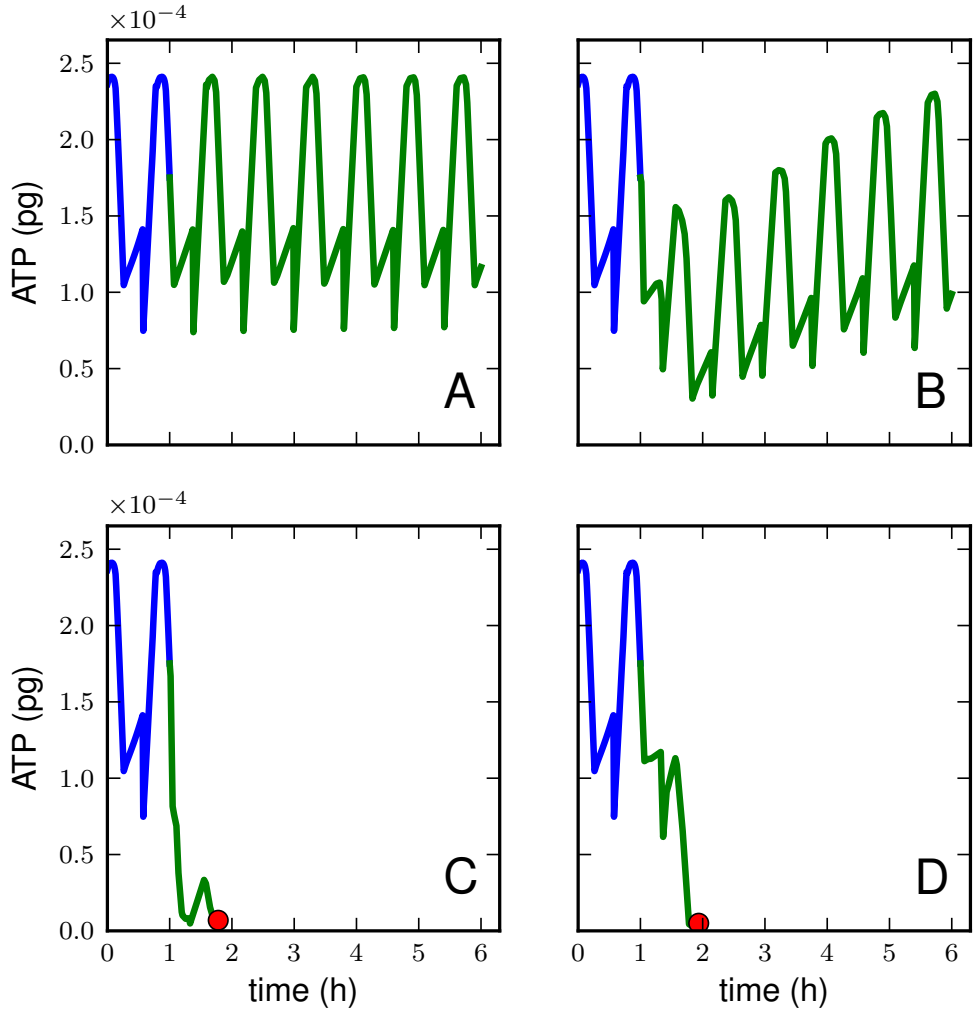


Figure 5.3: Effect of Pgi manipulations on adenosine triphosphate (ATP) mass and cell viability. Blue trajectories are the unaltered ATP mass over time, while the green trajectories represent the ATP mass after changes A-D. Red dots represent the time and state of cell death. A - Default trajectory. B - 25% reduction in Pgi mass, instantaneous. C - 60% reduction in Pgi mass, instantaneous. D - Total knockout of *pgi* gene, permanent.

of cell division, when the mass of every chemical species in the model is instantaneously halved. After cell division occurs, the mass gradually increases until the initiation of chromosome replication, when the synthesis of DNA and increased demand for RNA precursors causes a rapid consumption of ATP. Finally, when chromosome replication terminates, the ATP consumption rate decreases and we again observe a net increase in ATP mass until the cell division event occurs and it is once again halved.

Figures 5.3-B, C, and D demonstrate the effects of interventions related to the Pgi reaction. The cell can recover from certain reductions in this enzyme's activity (e.g. a temporary, step-change reduction in Pgi levels by 25%, Figure 5.3-B), while more drastic reductions (a 60% reduction in Pgi mass, or knocking out the *pgi* gene completely, Figure 5.3-C,D) result in cell death. Notably, the ATP level drops more rapidly in Figure 5.3-C than in Figure 5.3-D. This is because the intervention acts immediately on the protein Pgi, while the knockout mutation in 5.3-D acts upstream on the expression of *pgi*. Although the knockout ultimately has the same effect, its influence is slightly delayed so that another cell division is allowed to be completed before the cell fails.

5.6 Competitive Inhibition of Nutrient Uptake

The MCM connects the physiology of the minimal cell directly to its environment. The MCM could be used to guide development of appropriate nutrient media for synthetic cells. Except for inorganic ions, which are not tracked in the MCM, removing any of the external nutrients listed in Tables D.1 and D.2 causes the cell to fail. To further study the effect of environmental

nutrient modifications, model cells growing at steady-state were exposed to step-changes in the external concentrations of arginine, a competitive inhibitor of transport for other amino acids. Transport systems with multiple substrates are subject to competitive inhibition (Cheng and Prusoff, 1973). To reduce the total number of genes as much as possible, several transporters with broad specificity were included in the MCM. For example, the Bgt transport system, and ATP-Binding-Cassette (ABC) dimer found in *Synechocystis sp.*, is known to transport alanine, glutamine, glycine, leucine, proline, and serine (Quintero et al., 2001). The MCM accounts for multiple substrate inhibition using Michaelis-Menten competitive inhibition terms. Each transport rate law has one inhibition term for each alternative substrate, as described in Section 4.13.1. For example, a transporter that carries four substrates will have three external inhibition multipliers for each of its transport rate laws. This means that the concentrations of some substances cannot be arbitrarily increased because at some level they inhibit growth by causing the cell to be starved of another nutrient.

To exemplify the effect of competitive substrate inhibition on the viability of the MCM, the external concentration of arginine was increased 2x, 5x, and 10x (Figure 5.4). Arginine is transported into the cell by the Nat transport system of *Synechocystis* sp., which also transports histidine and lysine (Quintero et al., 2001). The rate of lysine uptake is described in Equations 5.3 and 5.4.

$$R_{Lys} = v_{R\text{-}Lys} \cdot K_{sat\text{-}Lys\text{-}ext} \cdot K_{sat\text{-}ATP} \cdot K_{i\text{-}Lys} \cdot K_{i\text{-}R\text{-}Lys} \cdot T_{Nat} \tag{5.3}$$

$$K_{i\text{-}R\text{-}Lys} = \frac{K_{i\text{-}R\text{-}Lys\text{-}Arg\text{-}ext}}{K_{i\text{-}R\text{-}Lys\text{-}Arg\text{-}ext} + Arg_{ext}} \cdot \frac{K_{i\text{-}R\text{-}Lys\text{-}His\text{-}ext}}{K_{i\text{-}R\text{-}Lys\text{-}His\text{-}ext} + His_{ext}} \tag{5.4}$$

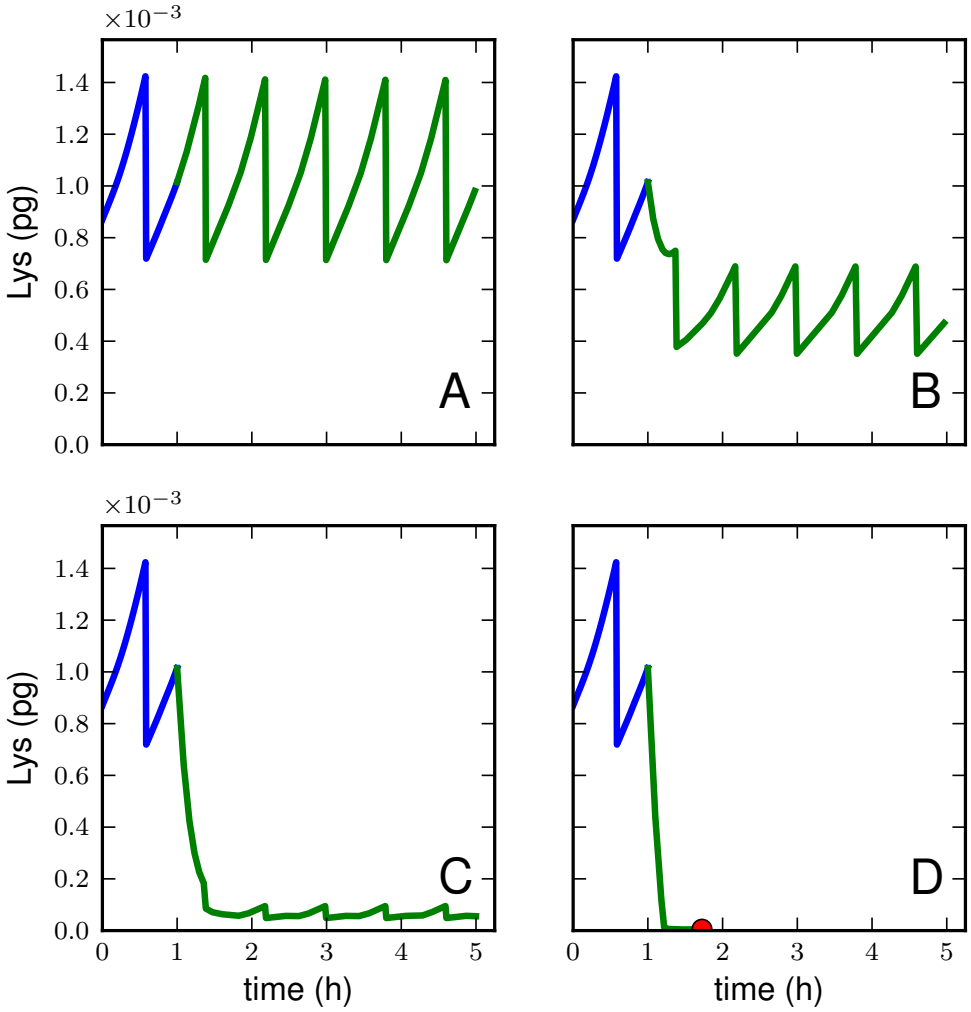


Figure 5.4: Effect of amino acid inhibition on lysine (Lys) mass and cell viability in response to increases in extracellular arginine (Arg). Blue trajectories are the unaltered lysine mass over time, while the green trajectories represent the lysine mass after changes A-D. Red dots represent the time and state of cell death. A - Default trajectory. B - 5x increase in the external concentration of arginine. C - 10x increase in the external concentration of arginine. D - 15x increase in the external concentration of arginine.

In Equation 5.3, R_{Lys} describes the rate of lysine uptake ($\frac{\text{pg}}{\text{h}}$), $v_{R\text{-}Lys}$ is the rate constant for lysine uptake ($\frac{\text{pg Lys}}{\text{h}\cdot\text{pg T}_{\text{Nat}}}$), $K_{sat\text{-}Lys\text{-}ext}$ and $K_{sat\text{-}ATP}$ are dimensionless Michaelis-Menten saturation terms for external lysine and cellular ATP, respectively, $K_{i\text{-}Lys}$ is a dimensionless Michaelis-Menten product inhibition term cellular lysine, $K_{i\text{-}R\text{-}Lys}$ is a dimensionless competitive inhibition term defined in Equation 5.4, and T_{Nat} is the mass of transporter T_{Nat} (pg). In Equation 5.4, $K_{i\text{-}R\text{-}Lys\text{-}Arg\text{-}ext}$ and $K_{i\text{-}R\text{-}Lys\text{-}His\text{-}ext}$ are inhibition constants ($\frac{\text{gm}}{\text{mL}}$) that describe transport inhibition by arginine and histidine, respectively on the lysine transport reaction.

Based on these equations, it is expected that the transport rate for lysine will drop as either arginine or histidine is increased in the medium. Figure 5.4 demonstrates such an effect, with lysine values becoming inhibitory somewhere between the 10x and 15x increase of the default concentration (Figure 5.4-C,D). This shows that there is an intermediate transition nutrient concentration where the cell transitions between life and death.

5.7 Comparison to Previous Work

There has been limited mathematical analysis of the minimal gene set proposed by Gil et al. (Gil et al., 2004; Gabaldón et al., 2007). A structural analysis revealed that a particular activity of nucleoside diphosphate kinase (Ndk) is not necessary for cells to achieve a steady-state (Gabaldón et al., 2007). Specifically, it was found that when the CTP + ADP $\leftrightarrow$ CDP + ATP activity (the NDK5 activity) was removed from the reaction network, the cell model could still find a steady-state.

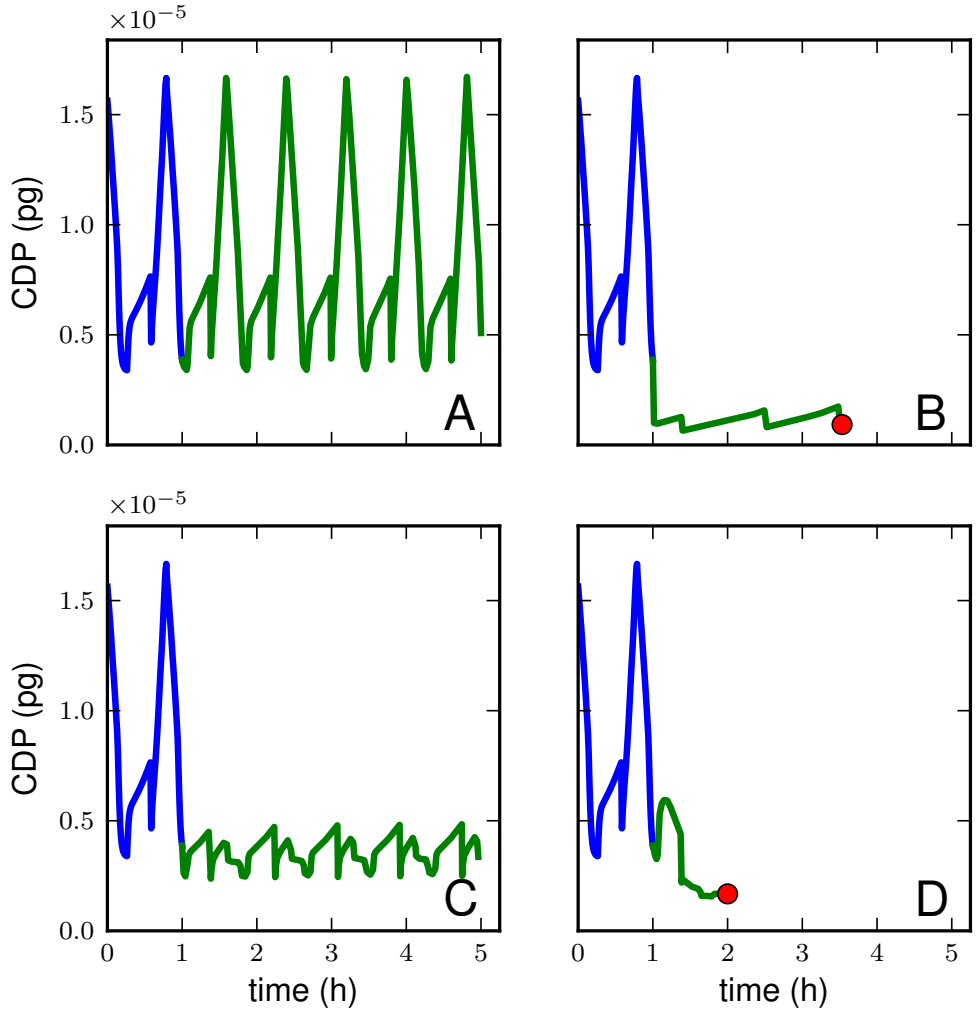


Figure 5.5: Effect of removing a particular activity of Ndk on cytidine diphosphate (CDP) mass and cell viability. Blue trajectories are the unaltered ATP mass over time, while the green trajectories represent the ATP mass after changes A-D. Red dots represent the time and state of cell death. A - Default trajectory. B - Removal of NDK5 activity. C - Permanent 25% reduction in v_{NDK5}. D - Total knockout of *ndk* gene.

To compare the MCM to those results, we performed interventions related to NDK5 activity (Figure 5.5). Figure 5.5-A shows the steady-state behavior of CDP mass (pg) with no intervention. In contrast to the result of Gabaldón et al. (2007), the current analysis shows that the NDK5 activity is necessary for cell survival. Specifically it is found that removing the NDK5 activity (Figure 5.5-B), a 25% reduction in the rate constant for the NDK5 reaction (Figure 5.5-C), and a total knockout of the *ndk* gene resulted in cell death. The discrepancy highlights the difference between a structural (stoichiometric) analysis of a metabolic network and a dynamic, whole-cell model. The approach used by Gabaldón et al. (2007) is limited to a subset of the minimal gene set in which cofactor metabolism was not considered. It is probable that the essentiality of the NDK5 activity is only revealed when the full network (i.e., a whole-cell model) is considered. Alternatively, the difference could be due to the differences in the interpretation of which reactions are reversible in the minimal cell. The MCM treats most reactions as irreversible or only weakly reversible, whereas the stoichiometric analysis by Gabaldón et al. (2007) considers many reactions to be fully reversible. Allowing more reversible reactions may provide the cell access to steady-states that are not possible in the MCM, and it may prove necessary for models of bacteria living in more complicated environments or with more diverse metabolism.

5.8 Conclusions

The ultimate goal of computational systems biology is to be able to ask a computer simulation any question that can be asked of *in vivo* models. The key to realizing that goal is the addition of mechanistic, chemical, and genomic

detail to models of whole-cells that are actively growing and dividing. In this chapter, a variety of computational experiments are presented that motivate further exploration of detailed hybrid bacterial cell models. It has been shown for the first time that it is possible to simulate a whole-cell whose behavior depends on its (i) metabolic rates and chemical state, (ii) genome in terms of expression of various genes, (iii) environment both in terms of direct nutrient starvation and competitive inhibition leading to starvation, and (iv) genomic sequence in terms of the locations of genes on the chromosome. The specific genetic manipulations discussed include knockouts for the *pgi* and *ndk* genes, as well as the variations of the position of a hypothetical gene insert. The application of phase plane analysis to the MCM has been demonstrated. An analysis of the MCM's response to an increase in arginine, which acts as a competitive inhibitor of the uptake of other amino acids, has also been presented. All of these behaviors are exhibited by a single-cell model that makes reasonable assumptions about cellular biochemistry, reaction rates, gene expression, and the effect of discrete physiological events on the cell's behavior. Therefore, the MCM makes substantial progress toward the computational systems biology's aims.

This type of computational experiment could have beneficial applications in synthetic biology. For example, the J. Craig Venter Institute has been actively pursuing the goal of synthesizing a cell with a small genome. They successfully transplanted a complete *Mycoplasma mycoides* chromosome into a *Mycoplasma capricolum* cell whose own genome had been removed (Lartigue et al., 2007). Next, they constructed a synthetic *Mycoplasma genitalium* genome (Gibson et al., 2008). Finally, they took the entire genome from *M. mycoides*, modified it in yeast using yeast genetic systems, and then transplanted the modified chromosome

into *M. capricolum* (Lartigue et al., 2009). This puts them very close to their ultimate goal of taking a wholly synthetic chromosome and using it as the starting genetic information for a new cell line. However, the project has taken longer than originally projected (Zimmer, 2003), and it is possible that part of the difficulty lies in finding an appropriate initial condition for the synthetic cell. The MCM could aid parallel efforts in synthetic biology by providing a framework to test the viability of cells with particular initial conditions.

REFERENCES

Bremer, D. P., H. (1996). *Modulation of Chemical Composition and other Parameters of the Cell by Growth Rate, in Escherichia coli and Salmonella: Cellular and Molecular Biology, F.C. Neidhart, Editor*. ASM Press.

Browning, S. T. and Shuler, M. L. (2001). Towards the development of a minimal cell model by generalization of a model of *Escherichia coli*: Use of dimensionless rate parameters. *Biotechnology and Bioengineering*, 76(3), 187–192.

Castellanos, M., Wilson, D. B., and Shuler, M. L. (2004). A modular minimal cell model: purine and pyrimidine transport and metabolism. *Proceedings of the National Academy of Sciences of the United States of America*, 101(17), 6681–6686. doi:10.1073/pnas.0400962101.

Cheng, Y. and Prusoff, W. H. (1973). Relationship between the inhibition constant (K1) and the concentration of inhibitor which causes 50 per cent inhibition (I50) of an enzymatic reaction. *Biochemical Pharmacology*, 22(23), 3099–3108.

Dandekar, T., Snel, B., Huynen, M., and Bork, P. (1998). Conservation of gene order: a fingerprint of proteins that physically interact. *Trends in Biochemical Sciences*, 23(9), 324–328.

Domach, M. M. (1983). *Refinement and Use of a Structured Model of a Single Cell of Escherichia coli for the Description of Ammonia-Limited Growth and Asynchonous Population Dynamics*. Ph.D. thesis, Cornell University.

Domach, M. M., Leung, S. K., Cahn, R. E., Cocks, G. G., and Shuler, M. L. (1984). Computer model for glucose-limited growth of a single cell of *Escherichia*

coli B/r-A. *Biotechnology and Bioengineering*, 26(9), 1140. doi:10.1002/bit.260260925.

Echtenkamp, P. L., Wilson, D. B., and Shuler, M. L. (2009). Cell cycle progression in *Escherichia coli* B/r affects transcription of certain genes: Implications for synthetic genome design. *Biotechnology and Bioengineering*, 102(3), 902–909. doi:10.1002/bit.22098.

Foley, P. L. and Shuler, M. L. (2010). Considerations for the design and construction of a synthetic platform cell for biotechnological applications. *Biotechnology and Bioengineering*, 105(1), 26–36. doi:10.1002/bit.22575.

Gabaldón, T., Peretó, J., Montero, F., Gil, R., Latorre, A., et al. (2007). Structural analyses of a hypothetical minimal metabolism. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 362(1486), 1751–1762. doi:10.1098/rstb.2007.2067.

Gibson, D. G., Benders, G. A., Andrews-Pfannkoch, C., Denisova, E. A., Baden-Tillson, H., et al. (2008). Complete chemical synthesis, assembly, and cloning of a *Mycoplasma genitalium* genome. *Science*, 319(5867), 1215–1220. doi:10.1126/science.1151721.

Gil, R., Silva, F. J., Pereto, J., and Moya, A. (2004). Determination of the core of a minimal bacterial gene set. *Microbiology and Molecular Biology Reviews*, 68(3), 518–537.

Kanehisa, M. and Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1), 27–30.

Lartigue, C., Glass, J. I., Alperovich, N., Pieper, R., Parmar, P. P., et al. (2007).

Genome transplantation in bacteria: changing one species to another. *Science*, 317(5838), 632–638. doi:10.1126/science.1144622.

Lartigue, C., Vashee, S., Algire, M. A., Chuang, R.-Y., Benders, G. A., et al. (2009). Creating bacterial strains from genomes that have been cloned and engineered in yeast. *Science*, 325(5948), 1693–1696. doi:10.1126/science.1173759.

May, R. M. (1972). Limit cycles in predator-prey communities. *Science*, 177(4052), 900–902. doi:10.1126/science.177.4052.900.

Mushegian, A. R. and Koonin, E. V. (1996). Gene order is not conserved in bacterial evolution. *Trends in Genetics*, 12(8), 289–290.

Quintero, M. J., Montesinos, M. L., Herrero, A., and Flores, E. (2001). Identification of genes encoding amino acid permeases by inactivation of selected ORFs from the synechocystis genomic sequence. *Genome Research*, 11(12), 2034–2040.

Seto, S. and Miyata, M. (1998). Cell reproduction and morphological changes in *Mycoplasma capricolum*. *Journal of Bacteriology*, 180(2), 256–264.

Tamames, J. (2001). Evolution of gene order conservation in prokaryotes. *Genome Biology*, 2(6), research0020.10020.

Tamames, J., Gonzlez-Moreno, M., Mingorance, J., Valencia, A., and Vicente, M. (2001). Bringing gene order into bacterial shape. *Trends in Genetics*, 17(3), 124–126.

Zimmer, C. (2003). Genomics - Tinker, tailor: Can Venter stitch together a genome from scratch? *Science*, 299(5609), 1006–1007.

CHAPTER 6

CONCLUSIONS AND RECOMMENDATIONS

6.1 Conclusions

The design of a chemically and genomically detailed Minimal Cell Model (MCM) is a major step in support of synthetic and systems biology. The overall goals of this research project were (i) to develop more powerful and flexible computational techniques for analysis of coarse-grained bacterial cell models, and (ii) to develop a model of a hypothetical bacterium with the minimum number of genes necessary and sufficient to support sustained division (i.e. an MCM). These goals were intentionally codependent. More flexible, object-oriented, and extensible computational techniques allowed the pursuit of a mathematical modeling framework of novel complexity and detail. At the same time, as development of the MCM progressed, it became clear that the underlying modeling methods needed to be updated to accommodate the expanded lists of genes, chemical species, and reactions. That need motivated development of new computational methods and modeling structures.

There is an ongoing effort to define a minimal gene set for prokaryotic life (Gil et al., 2004; Moya et al., 2009); however, there is currently no accepted method for testing the plausibility a minimal gene set once it is proposed. This dissertation has shown for the first time that it is possible to test the plausibility of a minimal gene set using a mathematical model of a whole chemoheterotrophic bacterial cell. The model cell is able to sustain growth and replication indefinitely in its optimally supportive culture environment. The chemically detailed nature of the model allows it to address sophisticated

experimental questions. Furthermore, as a tool of computational biology, it lends itself to being a building block for arbitrarily complex systems studies. For example, many cells could be simulated in parallel in an effort to simulate a bacterial cell culture, as in Domach and Shuler (1984a).

This dissertation describes the development of the MCM as well as its applications. Chapter 1 presents important considerations for making a mathematical model of a minimal cell. The motivation for developing mathematical models of bacterial cells, and in particular for developing an MCM, is discussed. Past work in computer modeling of bacteria is presented, and the concepts of minimal cells and minimal gene sets are introduced. Finally, the MCM is defined with reference to previous work on the project.

Chapter 2 introduces a sensitivity analysis method for hybrid bacterial cell models. While this method was ultimately not applied to the MCM, its development drove the redefinition of the Cornell *Escherichia coli* model in both MATLAB® and Systems Biology Markup Language (SBML) formats. The sensitivity analysis provides a method to identify particular submodules as prime candidates for delumping in hybrid cell models. A method for stability analysis is also presented, and it is demonstrated that the *E. coli* model has potential (via a Hopf bifurcation) for modulated quasi-periodic oscillations with a period larger than the doubling time of the cell. What this indicates is that some features of the model are not constant from generation to generation; rather, they repeat every two or more generations (Nikolaev et al., 2006). This is a complex, system-level outcome that is a direct result of the whole-cell hybrid approach.

Chapter 3 explains an updated version of the Cornell *E. coli* model that

links detailed genomic information about the location of *dnaA* genes and DnaA binding sites on the chromosome to physiological predictions. This is the first example of including detailed genomic information in a hybrid bacterial cell model; it lays the essential computational groundwork for the massive inclusion of new genes in the MCM. The model also suggests that the concentration of DNA binding boxes on the chromosome is critical to determining cell growth and behavior.

Chapter 4 describes the modeling structures used to create the Minimal Cell Model (MCM) as well as the submodels of metabolism and physiological processes that drive it. The MCM itself is the most significant outcome of this dissertation. We show for the first time that it is possible to test the hypotheses behind a minimal gene set using a chemically detailed, dynamic, whole-cell modeling approach. An MCM with 241 product-coding genes (those which produce protein or stable RNA products) is presented. This is supplementary to the minimal gene set proposed by (Gil et al., 2004). It is proposed that this set is genomically complete and codes for all the functions that a minimal chemoheterotrophic bacterium would require for sustained growth and division. The hybrid cell modeling approach originally used for a coarse-grained model of *E. coli* (Nikolaev et al., 2005) has been refined and made more rigorous for use with the MCM. As computational resources become faster and less expensive, larger systems should be tractable using these methods.

The variety of computational experiments presented in Chapter 5 motivate further exploration of detailed hybrid bacterial cell models. In particular, we show that it is possible to simulate a whole-cell whose behavior depends on its (i) metabolic rates and chemical state, (ii) genome in terms of expression of

various genes, (iii) environment both in terms of direct nutrient starvation and competitive inhibition leading to starvation, and (iv) genomic sequence in terms of the locations of genes on the chromosome. The specific genetic manipulations discussed in Chapter 5 include knockouts for the *pgi* and *ndk* genes, as well as the variation of the position of a hypothetical gene insert, *ins*. The application of phase plane analysis to the MCM has been demonstrated. An analysis of the MCM's response to an increase in arginine, which acts as a competitive inhibitor of the uptake of other amino acids, has also been presented. Previous work proposed that the so-called NDK5 activity of the *cmk* gene is not necessary for a minimal cell based on the Gil et al. (2004) gene set to survive (Gabaldón et al., 2007). The results presented here show that the NDK5 activity is essential, and that this essentiality is only revealed in the context of a whole-cell analysis like the MCM . All of these behaviors are exhibited by a single-cell model that makes reasonable assumptions about cellular biochemistry, reaction rates, gene expression, and the effect of discrete physiological events on the cell's behavior. By connecting biochemistry to physiological behavior, the MCM makes substantial progress toward the overall aims of computational systems biology.

The Shuler group has expertise in bacterial cell models that include the effects of discrete physiological events. These events depend on the chemical and genomic detail contained in the model, resulting in a clear connection between genomic sequence and physiological processes including DNA replication, transcription, translation, and cell division. The metabolic and transport reactions included in the MCM have been described. Detailed illustrations of the metabolic pathways in the MCM are included in Appendix C. The MCM is in balance with respect to redox potential and carbon flow, at least

to a reasonable approximation. Overall, we consider it to be a physiologically complete, chemically and genomically detailed representation of a minimal cell.

The model presented here is not the only possible minimal cell model. There is no evidence that there is one particular minimal cell (Gil et al., 2004). The current Minimal Cell Model has been established using the (Gil et al., 2004) minimal gene set. At 0.86 h^{-1}, the growth rate (μ_g) of the MCM simulated here is faster than one might expect for a minimal cell. However, it is proposed here that the absolute value of the growth rate is not critical. In some sense this growth rate is arbitrary. What is more important is the values of parameters *relative* to one another within a parameter set (Browning and Shuler, 2001). This dissertation establishes that it is possible to establish a minimal cell model using a coarse-grained approach. The MCM is, however, not unique. Other minimal gene sets could produce viable cells, just as there are alternate parameter sets that could drive the current model to a steady-state. It is precisely this ambiguity that motivates the development of computational methods for discriminating amongst minimal gene sets.

6.2 Recommended Project Extensions

There are significant portions of the new MCM that, while they are chemically detailed, still lack mechanistic detail, particularly when the physical structure of the cell must be recognized explicitly. Overall, the *chemical* detail present in the MCM will provide the ability to ask questions at a resolution that has, thus far, not been present in bacterial cell models. Adding *mechanistic* detail for a system of interest, however, could increase the value of the model for

some experimenters. Reasonable extensions to the base MCM presented in this dissertation are outlined below.

The formation of the cell septum during division in the MCM is catalyzed by the FtsZ protein, but there is no specific mechanism in the model that specifies that the septum should form precisely at the midcell. The current model assumes that septum formation occurs because of the influence of the FtsZ protein, without providing a mechanism for that behavior. Providing a mechanism for this split would allow the model to test whether the assumption of a 50/50 split at division is important. It would be possible to update the MCM with physical constraints that force this mode of division as in Surovtsev et al. (2009). The assumption of division at the midcell could also be relaxed, and the effects of asynchronous division could be investigated as in Domach and Shuler (1984b).

Another area lacking mechanistic detail is ribosome synthesis. The proposed minimal gene set includes 50 genes coding for ribosomal proteins (Gil et al., 2004). Those 50 genes have been included in the MCM, but no model for ribosome assembly is included, and the process is assumed to occur according to a Michaelis-Menten like rate. Although self-assembly of the small and large ribosomal subunits has been studied extensively (Culver, 2003; Talkington et al., 2005; Röhl and Nierhaus, 1982), there is not currently a mathematical model for the process that would be amenable to inclusion in the MCM. The protein and RNA components of the ribosome are all explicitly included in the MCM, so although ribosome formation is extremely complex, the MCM could be an ideal platform for testing hypotheses about how the ribosome forms. If an appropriate ribosomal assembly model were developed, including that model

in the MCM would be a logical extension to this work.

The genes on the MCM chromosome have been ordered somewhat arbitrarily. Research has shown that gene order on the chromosome is generally not conserved across long evolutionary distances in bacteria (Mushegian and Koonin, 1996; Dandekar et al., 1998; Tamames, 2001). Where gene order is conserved, it is often between pairs of genes whose protein products physically interact with each other (Dandekar et al., 1998), a concept which is captured in spirit by our use of gene clusters. The position of a gene can influence its expression via the gene dosage, both in nature and in the MCM (Foley and Shuler, 2010), and it would be beneficial to define a rationale for how genes should be ordered on a minimal chromosome.

It is desirable to have a glucose-controlled model so that the simulator can measure how different cellular attributes vary with growth rate. The MCM does not currently exhibit growth rate control via manipulation of external concentrations. Rather, the growth rate stays nearly constant as the glucose concentration is lowered until at some threshold glucose level the cell undergoes a very sharp transition to cell death. This shows that the model displays a high sensitivity to changes in its environment. Unlike real bacteria, a minimal cell has no alternative pathways to start when one pathway is shut down. In practice it would be necessary to have some way to control the cell's growth rate, and this is a good example of why it is important to draw a distinction between a minimal cell and a synthetic platform cell for biotechnology (Foley and Shuler, 2010). Cells intended for biotechnological applications must exhibit growth rate control through external nutrient manipulation, perhaps by including a stringent response

mechanism (Barker et al., 2001; Chatterji and Ojha, 2001). It should be noted that the stringent response is included in the standard Cornell *E. coli* model. Thus, methods to accomplish this extension are available, although more than the stringent response may be necessary to obtain adequate growth rate control.

It would be beneficial to have a more precise determination of the initial condition of the MCM using more detailed chemical composition measurements. The rate constant estimation procedure presented in Section 4.7.3 uses the initial mass of each chemical species in the cell to determine the required synthesis rates. Therefore, these rate constants are highly dependent on the initial concentrations of chemicals in the model. As the initial concentrations are refined, more meaning can be attributed to the rate constant estimates. Because the MCM is considered to represent a generalized chemoheterotrophic bacterium, the most useful data for each chemical would be the mass fraction for all chemical species present in a chemoheterotrophic bacterial cell.

There are a number of interesting extensions possible for the MCM's DNA synthesis module. The current model for DNA synthesis dictates that dNTP species are consumed with an average stoichiometry determined by the sequence of the genome. To our knowledge, this is the first model of a whole-cell that connects the consumption of DNA precursors to the DNA synthesis rate using explicit, genomic information. However, the model could go further by explicitly linking the stoichiometry of DNA synthesis to the fork position, which would provide a more accurate picture of how dNTPs are consumed over time. In a parallel update, one could also consider the effect of a more gradual change the cell's response to a step-change in gene dosage. Changes in gene dosage

when chromosome replication copies a gene are currently described with a step-function. In practice the ability of RNA polymerase to transcribe new gene copies may increase more gradually, and perhaps not even monotonically. It would be worthwhile to measure the importance of a particular gene dosage response in the MCM.

One could also add directionality to the chromosome representation. The current MCM labels the computer chromosome from position 0.0 at *Ori* to 1.0 at the terminus of replication. There is no distinction made as to which side of the chromosome a gene lies on, or which direction is the sense/antisense direction of a gene. Including this information would be the first step toward making the stoichiometry of DNA replication depend on the fork position of the DNA.

There are certain genes from the minimal gene set proposed by Gil et al. (2004) that are included in the MCM for completeness, but that have no mathematical connection to rates outside of the production of their own mRNA and protein products. For example, the genes necessary for DNA repair are included in the MCM, but their product concentrations do not directly influence the simulation behavior. By including a mechanism that would directly require these genes, the fidelity of the MCM to a hypothetical minimal cell could be increased, and the DNA repair machinery could have an explicit connection to cell survival. A future model release could include either random or averaged DNA damage mechanisms via specialized reactions.

There is strong interest in connecting more detailed physical chemistry to bacterial cell processes using bacterial cell models. The MCM could serve as a platform for this type of research. For example, the current MCM exists in an idealized environment with constant temperature. However, it would be

possible to study how temperature perturbations affect the cell behavior using Arrhenius expressions (Ataai and Shuler, 1986). Another important question related to physical chemistry is how the lipid composition of the cell membrane affects cell physiology. Using the MCM as a basis, a more realistic membrane with multiple lipid components could be introduced. The lipid composition of the membrane could be connected mathematically to simulation output using events. For example, for a lipid membrane with two lipid types, L1 and L2, the model could have events that are triggered when the ratio L1/L2 passes some threshold value. Such events could modulate cellular processes such as diffusion across the membrane.

From a synthetic biology perspective, it would be beneficial to use the MCM to determine what novel functions are necessary to help a cell survive challenges in the environment. One could propose a set of mutations to the cell that impart particular gains of function, and then automatically test to see which mutations allow the cell to overcome particular challenges. This experiment is not possible with the version of the MCM described in this dissertation, but the MCM does act as a step on the way to testing questions related to cell evolution.

The model is currently available in the Systems Biology Markup Language (SBML) format, which should make it accessible to a wider audience in computational biology. Currently, we can provide files and simulation tools for other research groups to work with the MCM. However, it still takes some technical expertise to download and make use of an SBML file. It would be advantageous to develop a web-based tool where a researcher could manipulate the model using the Internet. The primary challenge to making this useful is that the simulation takes a long time, and web-based tools are not usually

efficient for long simulations. An exciting possibility for sharing this model and generating new results would be to develop a site with a gallery of interesting results from the MCM, providing motivation for researchers to download and install the whole simulation package.

REFERENCES

Ataai, M. M. and Shuler, M. L. (1986). Mathematical-model for the control of ColE1 type plasmid replication. *Plasmid*, 16(3), 204–212.

Barker, M. M., Gaal, T., Josaitis, C. A., and Gourse, R. L. (2001). Mechanism of regulation of transcription initiation by ppGpp. I. effects of ppGpp on transcription initiation *in vivo* and *in vitro*. *Journal of Molecular Biology*, 305(4), 673–688.

Browning, S. T. and Shuler, M. L. (2001). Towards the development of a minimal cell model by generalization of a model of *Escherichia coli*: Use of dimensionless rate parameters. *Biotechnology and Bioengineering*, 76(3), 187–192.

Chatterji, D. and Ojha, A. K. (2001). Revisiting the stringent response, ppGpp and starvation signaling. *Current Opinion in Microbiology*, 4(2), 160–165.

Culver, G. M. (2003). Assembly of the 30S ribosomal subunit. *Biopolymers*, 68(2), 234–249. doi:10.1002/bip.10221.

Dandekar, T., Snel, B., Huynen, M., and Bork, P. (1998). Conservation of gene order: a fingerprint of proteins that physically interact. *Trends in Biochemical Sciences*, 23(9), 324–328.

Domach, M. M. and Shuler, M. L. (1984a). A finite representation model for an asynchronous culture of *Escherichia coli*. *Biotechnology and Bioengineering*, 26(8), 877–884.

Domach, M. M. and Shuler, M. L. (1984b). Testing of a potential mechanism for *Escherichia coli* temporal cycle imprecision with a structural model. *Journal of Theoretical Biology*, 106(4), 577–585.

Foley, P. L. and Shuler, M. L. (2010). Considerations for the design and construction of a synthetic platform cell for biotechnological applications. *Biotechnology and Bioengineering*, 105(1), 26–36. doi:10.1002/bit.22575.

Gabaldón, T., Peretó, J., Montero, F., Gil, R., Latorre, A., et al. (2007). Structural analyses of a hypothetical minimal metabolism. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 362(1486), 1751–1762. doi:10.1098/rstb.2007.2067.

Gil, R., Silva, F. J., Pereto, J., and Moya, A. (2004). Determination of the core of a minimal bacterial gene set. *Microbiology and Molecular Biology Reviews*, 68(3), 518–537.

Moya, A., Gil, R., Latorre, A., Peret, J., Garcilln-Barcia, M. P., et al. (2009). Toward minimal bacterial cells: evolution vs. design. *FEMS Microbiology Reviews*, 33(1), 225–235. doi:10.1111/j.1574-6976.2008.00151.x.

Mushegian, A. R. and Koonin, E. V. (1996). Gene order is not conserved in bacterial evolution. *Trends in Genetics*, 12(8), 289–290.

Nikolaev, E., Atlas, J., and Shuler, M. L. (2006). Computer models of bacterial cells: from generalized coarse-grained to genome-specific modular models. *Journal of Physics: Conference Series*, 46, 322–326.

Nikolaev, E. V., Burgard, A. P., and Maranas, C. D. (2005). Elucidation and structural analysis of conserved pools for genome-scale metabolic reconstructions. *Biophysical Journal*, 88(1), 37–49. doi:10.1529/biophysj.104.043489.

Röhl, R. and Nierhaus, K. H. (1982). Assembly map of the large subunit (50S) of

Escherichia coli ribosomes. *Proceedings of the National Academy of Sciences of the United States of America*, 79(3), 729–733.

Surovtsev, I. V., Zhang, Z., Lindahl, P. A., and Morgan, J. J. (2009). Mathematical modeling of a minimal protocell with coordinated growth and division. *Journal of Theoretical Biology*, 260(3), 422–429. doi:10.1016/j.jtbi.2009.06.001.

Talkington, M. W. T., Siuzdak, G., and Williamson, J. R. (2005). An assembly landscape for the 30S ribosomal subunit. *Nature*, 438(7068), 628–632. doi:10.1038/nature04261.

Tamames, J. (2001). Evolution of gene order conservation in prokaryotes. *Genome Biology*, 2(6), research0020.10020.

APPENDIX A

MODEL NAMING CONVENTIONS

A.1 Naming Conventions

The supplemental website for the Minimal Cell Model (MCM) will serve as a repository for all the model simulation code, structure definitions, and model module definitions (see Appendix I). This Appendix describes naming conventions used in the code and throughout the dissertation.

The MCM has 408 chemical species defined. In this dissertation, a species is generally referred to in *italic* font when referring to its mass, and in regular font when referring to the chemical itself. The species are usually named according to commonly accepted biochemical abbreviations, and when there is no common abbreviation the species is defined in the comments of the code.

Reactions are named according to whether they are considered as synthesis or degradation reactions. 'S' subscripts denote synthetic reactions, while 'D' subscripts denote degradation reactions. For example, $\mathrm{f6P_S}$ is the synthesis reaction for f6P (fructose-6P, or fructose-6-phosphate). The degradation subscript is only applied to degradation of macromolecules (e.g. mRNA or protein species). In this dissertation, reactions are generally referred to in *italic* font when referring to their quantitative rate, and in regular font when referring to the reaction itself.

The model simulation code automatically creates an assignment rule for the reaction rate of each reaction so that those rates can be referred to elsewhere in the model. For example, the rate of tryptophan export, R_{Trp}, is set as an

assignment rule so that its value can be used both in calculating tryptophan transport and in calculating the rate of ATP consumption related to the proton motive force loss used to import the molecule.

Reaction rate constants, saturation constants, and inhibition constants are all automatically named in the following patterns:

- For metabolic reactions that do not involve macromolecules, the reaction rate constants are named v_X, where X is the name of the reaction.
- For metabolic reactions involving the synthesis or degradation of mRNA or protein, reaction rate constants are named k_X, where X is the name of the macromolecule being synthesized or degraded.
- Saturation constants are named $Ks_{Y\text{-}Z}$ where Y is the name of the reaction being activated, and Z is the name of the activating chemical species. For example, $Ks_{f6P\text{-}S\text{-}g6P}$ is the saturation constant describing the effect of g6p (glucose-6P, or glucose-6-phosphate) on f6P (fructose-6P) synthesis.
- Inhibition constants are named $Ki_{Y\text{-}Z}$ where Y is the name of the reaction being inhibited, and Z is the name of the inhibiting chemical species. For example, $Ki_{R\text{-}Gln\text{-}Leu\text{-}ext}$ is the inhibition constant describing the effect of external leucine on glutamine transport.

A.2 Lumped Chemical Species

As described in Section 4.8.1, the MCM defines a number of lumped chemical species for convenience. For example, M_1 describes the total mass of all protein species in the model. These coarse-grained variables are inspired by the

previous modeling work in the Shuler group (Domach et al., 1984; Browning and Shuler, 2001; Castellanos et al., 2004, 2007). Even though the current model is much more chemically detailed, having access to the concepts of coarse-grained bacterial species modeling was critical for establishing the roles of gene clusters and their products. The mRNA and protein species associated with a particular gene cluster are considered as single mathematical entities in the MCM. Practically speaking, with the introduction of gene clusters and their products (Section 4.12) we have coarse-grained the action of particular groups of enzymes where the model lacks sufficient mechanistic details to distinguish their roles.

REFERENCES

Browning, S. T. and Shuler, M. L. (2001). Towards the development of a minimal cell model by generalization of a model of *Escherichia coli*: Use of dimensionless rate parameters. *Biotechnology and Bioengineering*, 76(3), 187–192.

Castellanos, M., Kushiro, K., Lai, S. K., and Shuler, M. L. (2007). A genomically/chemically complete module for synthesis of lipid membrane in a minimal cell. *Biotechnology and Bioengineering*, 97(2), 397–409. doi:10.1002/bit.21251.

Castellanos, M., Wilson, D. B., and Shuler, M. L. (2004). A modular minimal cell model: purine and pyrimidine transport and metabolism. *Proceedings of the National Academy of Sciences of the United States of America*, 101(17), 6681–6686. doi:10.1073/pnas.0400962101.

Domach, M. M., Leung, S. K., Cahn, R. E., Cocks, G. G., and Shuler, M. L. (1984). Computer model for glucose-limited growth of a single cell of *Escherichia coli* B/r-A. *Biotechnology and Bioengineering*, 26(9), 1140. doi:10.1002/bit.260260925.

APPENDIX B

MINIMAL GENE SET USED IN THE MINIMAL CELL MODEL

Our model implements a whole-cell dynamic model of a single cell that contains the minimal gene set described by Gil et al. (2004). The authors break their minimal gene set into five major categories:

1. Information Storage and Processing
2. Protein Processing, Folding, and Secretion
3. Cellular Processes
4. Energetic and Intermediate Metabolism
5. Poorly Characterized

The specifics of the minimal gene set used in the MCM, included differences with that proposed by Gil et al. (2004), are included in Section 4.20. Sequence information for each gene in the MCM was obtained from the KEGG database (Kanehisa and Goto, 2000), and Table B.1 summarizes how many genes came from each source organism. Table 4.8 in Chapter 4 shows a summary of how many genes fall into particular functional categories in the MCM. Table B.2 lists the genes from the (Gil et al., 2004) gene set that were not included in the MCM. Finally, a full listing of the genes in the MCM is presented in Table B.3.

Table B.1: Distribution of source genomes for finding sequences for the genes in the minimal gene set. *bpu - Bacillus pumilus*. *bsu - Bacillus subtilis*. *chu - Cytophaga hutchinsonii*. *eco - Escherichia coli*. *mge - Mycoplasma genitalium*. *rsp - Rhodobacter sphaeroides*. *syc - Synechococcus elongatus*. *wbr - Wigglesworthia brevipalpis*.

Organism	**Number of Genes**
bpu	1
bsu	10
chu	1
eco	59
mge	162
rsp	1
syc	4
wbr	3

Table B.2: Genes from the minimal gene set proposed by (Gil et al., 2004) that are excluded from the Minimal Cell Model.

Category	**Genes**
Protein Posttranslational Modification	*pepA*
Poorly Characterized	*mesJ*
	ybeY
	ycfF
	yoaE
	yqgF
	yraL

Table B.3: Minimal gene set used in the Minimal Cell Model. The genes are grouped by functional category. These categories correspond to the 'Subcategory' columns of Table 1 in Gil et al. (2004). The MCM Id refers to the gene's identifier in the MCM. The Gene column shows the gene, or genes in the case of a gene-cluster, that correspond to the given Id. Note that for gene clusters, a single Id is associated with several genes. The Species column lists which organism the sequence used in the MCM was obtained from. The Species abbreviations correspond to those listed in Table B.1. Function is a short description of the function of that gene. EC identifiers refer to the Enzyme Classification system.

Category	MCM Id	Gene(s)	Species	Function
Basic DNA replication machinery (14 genes)	$dnaB$	$dnaB$	*mge*	replicative DNA helicase
	$dnaG$	$dnaG$	*eco*	DNA primase (EC:2.7.7.-)
	$hupA$	$hupA$	*eco*	HU, DNA-binding transcriptional regulator, alpha subunit
	$mraW$	$mraW$	*mge*	S-adenosyl-methyltransferase MraW
	$replisome$	$dnaE$	*mge*	DNA primase
		$dnaN$	*mge*	DNA polymerase III, beta subunit (EC:2.7.7.7)
		$dnaQ$	*eco*	DNA polymerase III epsilon subunit (EC:2.7.7.7)
		$dnaX$	*eco*	DNA polymerase III/DNA elongation factor III, tau and gamma subunits (EC:2.7.7.7)
		$gyrA$	*mge*	DNA gyrase subunit A (EC:5.99.1.3)
		$gyrB$	*mge*	DNA gyrase subunit B (EC:5.99.1.3)

Table B.3 (Continued)

Category	MCM ID	Gene(s)	Species	Function
		$holA$	*eco*	DNA polymerase III, delta subunit (EC:2.7.7.7)
		$holB$	*eco*	DNA polymerase III, delta prime subunit (EC:2.7.7.7)
		lig	*wbr*	DNA ligase (NAD dependent)
		ssb	*mge*	single-strand binding protein family
Basic transcription machinery (8 genes)	pol_{RNA}	$deaD$	*mge*	DEAD-box ATP dependent DNA helicase
		$greA$	*mge*	transcription elongation factor GreA
		$nusA$	*mge*	transcription elongation factor NusA
		$nusG$	*eco*	transcription termination factor
		$rpoA$	*mge*	DNA-directed RNA polymerase subunit alpha (EC:2.7.7.6)
		$rpoB$	*mge*	DNA-directed RNA polymerase subunit beta (EC:2.7.7.6)
		$rpoC$	*mge*	DNA-directed RNA polymerase subunit beta′ (EC:2.7.7.6)
		$rpoD$	*mge*	RNA polymerase sigma factor

Table B.3 (Continued)

Category	MCM ID	Gene(s)	Species	Function
Biosynthesis of Cofactors (12 genes)	$coaA$	$coaA$	*eco*	pantothenate kinase (EC:2.7.1.33)
	$coaD$	$coaD$	*eco*	pantetheine-phosphate adenylyltransferase (EC:2.7.7.3)
	$coaE$	$coaE$	*eco*	dephospho-CoA kinase (EC:2.7.1.24)
	dfp	dfp	*eco*	fused 4′-phosphopantothenoylcysteine decarboxylase/phosphopantothenoylcysteine synthetase, FMN-binding (EC:4.1.1.36 6.3.2.5)
	$folA$	$folA$	*eco*	dihydrofolate reductase (EC:1.5.1.3)
	$glyA$	$glyA$	*mge*	serine hydroxymethyltransferase (EC:2.1.2.1)
	$metK$	$metK$	*eco*	S-adenosylmethionine synthetase (EC:2.5.1.6)
	$nadR$	$nadR$	*eco*	bifunctional DNA-binding transcriptional repressor/NMN adenylyltransferase
	$nadV$	$nadV$	*chu*	putative nicotinate phosphoribosyltransferase (EC:2.4.2.11)
	$pdxY$	$pdxY$	*eco*	pyridoxamine kinase (EC:2.7.1.35)
	$ribF$	$ribF$	*eco*	bifunctional riboflavin kinase/FAD synthetase (EC:2.7.1.26 2.7.7.2)

Table B.3 (Continued)

Category	MCM ID	Gene(s)	Species	Function
	yloS	*yloS*	*bpu*	thiamin pyrophosphokinase
Biosynthesis of nucleotides (15 genes)	*adk*	*adk*	*mge*	adenylate kinase (EC:2.7.4.3)
	dcd	*dcd*	*eco*	2′-deoxycytidine 5′-triphosphate deaminase (EC:3.5.4.13)
	gmk	*gmk*	*mge*	guanylate kinase (EC:2.7.4.8)
	hpt	*hpt*	*mge*	hypoxanthine phosphoribosyltransferase (EC:2.4.2.8)
	ndk	*ndk*	*eco*	multifunctional nucleoside diphosphate kinase and apyrimidinic endonuclease and 3′-phosphodiesterase (EC:2.7.4.6)
	nrdEF	*nrdE*	*mge*	ribonucleotide-diphosphate reductase subunit alpha (EC:1.17.4.1)
		nrdF	*mge*	ribonucleotide-diphosphate reductase subunit beta (EC:1.17.4.1)
	ppa	*ppa*	*mge*	inorganic pyrophosphatase (EC:3.6.1.1)
	prsA	*prsA*	*bsu*	molecular chaperone lipoprotein
	pyrG	*pyrG*	*eco*	CTP synthetase (EC:6.3.4.2)
	thyA	*thyA*	*mge*	thymidylate synthase (EC:2.1.1.45)
	tmk	*tmk*	*mge*	thymidylate kinase (EC:2.7.4.9)

Table B.3 (Continued)

Category	MCM ID	Gene(s)	Species	Function
	$trxA$	$trxA$	*mge*	thioredoxin
	$trxB$	$trxB$	*mge*	thioredoxin-disulfide reductase (EC:1.8.1.9)
	upp	upp	*mge*	uracil phosphoribosyltransferase (EC:2.4.2.9)
Cell division	$ftsZ$	$ftsZ$	*mge*	cell division protein FtsZ
DNA repair, restriction, and modificaiton (3 genes)	dna_{rep}	nth	*eco*	DNA glycosylase and apyrimidinic (AP) lyase (endonuclease III) (EC:4.2.99.18)
		$polA$	*eco*	fused DNA polymerase I 5′->3′ polymerase/3′->5′ exonuclease/5′->3′ exonuclease (EC:2.7.7.7)
		ung	*mge*	uracil-DNA glycosylase
Glycolysis (10 genes)	eno	eno	*mge*	phosphopyruvate hydratase (EC:4.2.1.11)
	$fbaA$	$fbaA$	*eco*	fructose-bisphosphate aldolase, class II (EC:4.1.2.13)

Table B.3 (Continued)

Category	MCM ID	Gene(s)	Species	Function
	$gapA$	$gapA$	*eco*	glyceraldehyde-3-phosphate dehydrogenase A (EC:1.2.1.12)
	$gpmA$	$gpmA$	*eco*	phosphoglyceromutase 1 (EC:5.4.2.1)
	ldh	ldh	*mge*	L-lactate dehydrogenase/malate dehydrogenase (EC:1.1.1.27)
	$pfkA$	$pfkA$	*mge*	6-phosphofructokinase (EC:2.7.1.11)
	pgi	pgi	*mge*	glucose-6-phosphate isomerase (EC:5.3.1.9)
	pgk	pgk	*mge*	phosphoglycerate kinase (EC:2.7.2.3)
	$pykA$	$pykA$	*eco*	pyruvate kinase II (EC:2.7.1.40)
	$tpiA$	$tpiA$	*mge*	triosephosphate isomerase (EC:5.3.1.1)
Lipid metabolism (7 genes)	$cdsA$	$cdsA$	*eco*	CDP-diglyceride synthase (EC:2.7.7.41)
	$fadD$	$fadD$	*eco*	acyl-CoA synthetase (long-chain-fatty-acid–CoA ligase) (EC:6.2.1.3)
	$gpsA$	$gpsA$	*eco*	glycerol-3-phosphate dehydrogenase (NAD+) (EC:1.1.1.94)
	$plsB$	$plsB$	*eco*	glycerol-3-phosphate O-acyltransferase (EC:2.3.1.15)

Table B.3 (Continued)

Category	MCM ID	Gene(s)	Species	Function
	$plsC$	$plsC$	*mge*	1-acyl-sn-glycerol-3-phosphate acyltransferase, putative (EC:2.3.1.51)
	psd	psd	*eco*	phosphatidylserine decarboxylase (EC:4.1.1.65)
	$pssA$	$pssA$	*eco*	phosphatidylserine synthase (CDP-diacylglycerol-serine O-phosphatidyltransferase) (EC:2.7.8.8)
Pentose phosphate pathway (4 genes)	$glpX$	$glpX$	*eco*	fructose 1,6-bisphosphatase II (EC:3.1.3.11)
	rpe	rpe	*eco*	D-ribulose-5-phosphate 3-epimerase (EC:5.1.3.1)
	$rpiA$	$rpiA$	*eco*	ribose 5-phosphate isomerase, constitutive (EC:5.3.1.6)
	tkt	tkt	*bsu*	transketolase (EC:2.2.1.1)
Protein folding (5 genes)	$prot_{fold}$	$dnaJ$	*mge*	chaperone protein DnaJ
		$dnaK$	*mge*	molecular chaperone DnaK
		$groEL$	*mge*	chaperonin GroEL
		$groES$	*mge*	co-chaperonin GroES

Table B.3 (Continued)

Category	MCM ID	Gene(s)	Species	Function
		$grpE$	*mge*	co-chaperone GrpE
Protein posttranslational modification (3 genes)	deg_{RNA}	pnp	*eco*	polynucleotide phosphorylase/polyadenylase (EC:2.7.7.8)
		rnc	*mge*	ribonuclease III (EC:3.1.26.3)
	map	map	*mge*	methionine aminopeptidase, type I (EC:3.4.11.18)
Protein translocation and secretion (5 genes)	$prot_{transloc}$	ffh	*mge*	signal recognition particle protein
		$ftsY$	*mge*	signal recognition particle-docking protein FtsY
		$secA$	*mge*	preprotein translocase subunit SecA
		$secE$	*mge*	preprotein translocase subunit SecE
		$secY$	*mge*	preprotein translocase subunit SecY
Protein turnover (3 genes)	deg_{M1}	gcp	*bsu*	putative DNA-binding/iron metalloprotein/AP endonuclease
		$hflB$	*wbr*	ATP-dependent protease

Table B.3 (Continued)

Category	MCM ID	Gene(s)	Species	Function
		lon	*mge*	ATP-dependent protease La (EC:3.4.21.53)
Proton motive force generation (9 genes)	*pmf*	*atpA*	*mge*	F0F1 ATP synthase subunit alpha (EC:3.6.3.14)
		atpB	*mge*	F0F1 ATP synthase subunit A (EC:3.6.3.14)
		atpC	*mge*	F0F1 ATP synthase subunit epsilon (EC:3.6.3.14)
		atpD	*mge*	F0F1 ATP synthase subunit beta (EC:3.6.3.14)
		atpE	*mge*	F0F1 ATP synthase subunit C (EC:3.6.3.14)
		atpF	*mge*	F0F1 ATP synthase subunit B (EC:3.6.3.14)
		atpG	*mge*	F0F1 ATP synthase subunit gamma (EC:3.6.3.14)
		atpH	*mge*	F0F1 ATP synthase subunit delta (EC:3.6.3.14)
		yidC	*eco*	membrane protein insertase

Table B.3 (Continued)

Category	MCM ID	Gene(s)	Species	Function
Ribosomal RNA (rRNA) (3 genes)	rti	$rrfA$	*mge*	5S ribosomal RNA
		$rrlA$	*mge*	23S ribosomal RNA
		$rrsA$	*mge*	16S ribosomal RNA
Transfer RNA (tRNA) (20 genes)	$tRNA_{Ala}$	$tRNA_{Ala}$	*mge*	tRNA for Ala
	$tRNA_{Arg}$	$tRNA_{Arg}$	*mge*	tRNA for Arg
	$tRNA_{Asn}$	$tRNA_{Asn}$	*mge*	tRNA for Asn
	$tRNA_{Asp}$	$tRNA_{Asp}$	*mge*	tRNA for Asp
	$tRNA_{Cys}$	$tRNA_{Cys}$	*mge*	tRNA for Cys
	$tRNA_{Gln}$	$tRNA_{Gln}$	*mge*	tRNA for Gln
	$tRNA_{Glu}$	$tRNA_{Glu}$	*mge*	tRNA for Glu
	$tRNA_{Gly}$	$tRNA_{Gly}$	*mge*	tRNA for Gly
	$tRNA_{His}$	$tRNA_{His}$	*mge*	tRNA for His
	$tRNA_{Ile}$	$tRNA_{Ile}$	*mge*	tRNA for Ile
	$tRNA_{Leu}$	$tRNA_{Leu}$	*mge*	tRNA for Leu
	$tRNA_{Lys}$	$tRNA_{Lys}$	*mge*	tRNA for Lys
	$tRNA_{Met}$	$tRNA_{Met}$	*mge*	tRNA for Met

Table B.3 (Continued)

Category	MCM ID	Gene(s)	Species	Function
	$tRNA_{Phe}$	$tRNA_{Phe}$	*mge*	tRNA for Phe
	$tRNA_{Pro}$	$tRNA_{Pro}$	*mge*	tRNA for Pro
	$tRNA_{Ser}$	$tRNA_{Ser}$	*mge*	tRNA for Ser
	$tRNA_{Thr}$	$tRNA_{Thr}$	*mge*	tRNA for Thr
	$tRNA_{Trp}$	$tRNA_{Trp}$	*bsu*	tRNA for Trp
	$tRNA_{Tyr}$	$tRNA_{Tyr}$	*mge*	tRNA for Tyr
	$tRNA_{Val}$	$tRNA_{Val}$	*mge*	tRNA for Val
Translation factors (12 genes)	$transF$	efp	*mge*	elongation factor P
		frr	*mge*	ribosome recycling factor
		$fusA$	*eco*	protein chain elongation factor EF-G, GTP-binding
		$hemK$	*wbr*	N5-glutamine methyltransferase, modulation of release factors activity
		$infA$	*mge*	translation initiation factor IF-1
		$infB$	*mge*	translation initiation factor IF-2
		$infC$	*mge*	translation initiation factor IF-3
		$lepA$	*mge*	GTP-binding protein LepA

Table B.3 (Continued)

Category	MCM ID	Gene(s)	Species	Function
		$prfA$	*mge*	peptide chain release factor 1
		$smpB$	*mge*	SsrA-binding protein
		tsf	*mge*	elongation factor Ts
		$tufA$	*eco*	protein chain elongation factor EF-Tu (duplicate of TufB)
Translation: aminoacyl-tRNA synthesis (21 genes)	$alaS$	$alaS$	*mge*	alanyl-tRNA synthetase (EC:6.1.1.7)
	$argS$	$argS$	*mge*	arginyl-tRNA synthetase (EC:6.1.1.19)
	$asnS$	$asnS$	*eco*	asparaginyl tRNA synthetase (EC:6.1.1.22)
	$aspS$	$aspS$	*mge*	aspartyl-tRNA synthetase (EC:6.1.1.12)
	$cysS$	$cysS$	*mge*	cysteinyl-tRNA synthetase (EC:6.1.1.16)
	$glnS$	$glnS$	*eco*	glutamyl-tRNA synthetase (EC:6.1.1.18)
	$gltX$	$gltX$	*mge*	glutamyl-tRNA synthetase (EC:6.1.1.17)
	$glyS$	$glyS$	*eco*	glycine tRNA synthetase, beta subunit (EC:6.1.1.14)
	$hisS$	$hisS$	*mge*	histidyl-tRNA synthetase (EC:6.1.1.21)
	$ileS$	$ileS$	*mge*	isoleucyl-tRNA synthetase (EC:6.1.1.5)

Table B.3 (Continued)

Category	MCM ID	Gene(s)	Species	Function
	$leuS$	$leuS$	*mge*	leucyl-tRNA synthetase (EC:6.1.1.4)
	$lysS$	$lysS$	*mge*	lysyl-tRNA synthetase (EC:6.1.1.6)
	$metS$	$metS$	*mge*	methionyl-tRNA synthetase (EC:6.1.1.10)
	$pheS$	$pheS$	*mge*	phenylalanyl-tRNA synthetase subunit alpha (EC:6.1.1.20)
		$pheT$	*mge*	phenylalanyl-tRNA synthetase subunit beta (EC:6.1.1.20)
	$proS$	$proS$	*mge*	prolyl-tRNA synthetase (EC:6.1.1.15)
	$serS$	$serS$	*mge*	seryl-tRNA synthetase (EC:6.1.1.11)
	$thrS$	$thrS$	*mge*	threonyl-tRNA synthetase (EC:6.1.1.3)
	$trpS$	$trpS$	*mge*	tryptophanyl-tRNA synthetase (EC:6.1.1.2)
	$tyrS$	$tyrS$	*mge*	tyrosyl-tRNA synthetase (EC:6.1.1.1)
	$valS$	$valS$	*mge*	valyl-tRNA synthetase (EC:6.1.1.9)
Translation: ribosomal proteins (50 genes)	$ribO$	$rplA$	*mge*	50S ribosomal protein L1
		$rplB$	*mge*	50S ribosomal protein L2
		$rplC$	*mge*	50S ribosomal protein L3
		$rplD$	*mge*	50S ribosomal protein L4

Table B.3 (Continued)

Category	MCM ID	Gene(s)	Species	Function
		$rplE$	*mge*	50S ribosomal protein L5
		$rplF$	*mge*	50S ribosomal protein L6
		$rplI$	*mge*	50S ribosomal protein L9
		$rplJ$	*mge*	50S ribosomal protein L10
		$rplK$	*mge*	50S ribosomal protein L11
		$rplL$	*mge*	50S ribosomal protein L7/L12
		$rplM$	*mge*	50S ribosomal protein L13
		$rplN$	*mge*	50S ribosomal protein L14
		$rplO$	*mge*	50S ribosomal protein L15
		$rplP$	*mge*	50S ribosomal protein L16
		$rplQ$	*mge*	50S ribosomal protein L17
		$rplR$	*mge*	50S ribosomal protein L18
		$rplS$	*mge*	50S ribosomal protein L19
		$rplT$	*mge*	50S ribosomal protein L20
		$rplU$	*mge*	50S ribosomal protein L21
		$rplV$	*mge*	50S ribosomal protein L22
		$rplW$	*mge*	50S ribosomal protein L23

Table B.3 (Continued)

Category	MCM ID	Gene(s)	Species	Function
		$rplX$	*mge*	50S ribosomal protein L24
		$rpmA$	*mge*	50S ribosomal protein L27
		$rpmB$	*mge*	50S ribosomal protein L28
		$rpmC$	*mge*	50S ribosomal protein L29
		$rpmE$	*mge*	50S ribosomal protein L31
		$rpmF$	*mge*	50S ribosomal protein L32
		$rpmG$	*mge*	50S ribosomal protein L33
		$rpmH$	*mge*	50S ribosomal protein L34
		$rpmI$	*mge*	50S ribosomal protein L35
		$rpmJ$	*mge*	50S ribosomal protein L36
		$rpsB$	*mge*	30S ribosomal protein S2
		$rpsC$	*mge*	30S ribosomal protein S3
		$rpsD$	*mge*	30S ribosomal protein S4
		$rpsE$	*mge*	30S ribosomal protein S5
		$rpsF$	*mge*	30S ribosomal protein S6
		$rpsG$	*eco*	30S ribosomal subunit protein S7
		$rpsH$	*mge*	30S ribosomal protein S8

Table B.3 (Continued)

Category	MCM ID	Gene(s)	Species	Function
		rpsI	*mge*	30S ribosomal protein S9
		rpsJ	*mge*	30S ribosomal protein S10
		rpsK	*eco*	30S ribosomal subunit protein S11
		rpsL	*mge*	30S ribosomal protein S12
		rpsM	*mge*	30S ribosomal protein S13
		rpsN	*mge*	30S ribosomal protein S14
		rpsO	*mge*	30S ribosomal protein S15
		rpsP	*mge*	30S ribosomal protein S16
		rpsQ	*mge*	30S ribosomal protein S17
		rpsR	*mge*	30S ribosomal protein S18
		rpsS	*mge*	30S ribosomal protein S19
		rpsT	*mge*	30S ribosomal protein S20
Translation: ribosome function, maturation, and modification (7 genes)	*ribM*	*cspR*	*bsu*	putative rRNA methylase (EC:2.1.1.-)
		engA	*mge*	GTP-binding protein EngA
		era	*mge*	GTP-binding protein Era
		ksgA	*mge*	dimethyladenosine transferase

Table B.3 (Continued)

Category	MCM ID	Gene(s)	Species	Function
		obg	*bsu*	
		$rbfA$	*eco*	30S ribosome binding factor
		$ychF$	*eco*	predicted GTP-binding protein
Translation: tRNA maturation and modificaiton (6 genes)	mat_{tRNA}	$iscS$	*eco*	cysteine desulfurase (tRNA sulfurtransferase), PLP-dependent (EC:2.8.1.7)
		$mnmA$	*mge*	tRNA (5-methylaminomethyl-2-thiouridylate)-methyltransferase (EC:2.1.1.61)
		$mnmE$	*eco*	GTPase
		$mnmG$	*eco*	5-methylaminomethyl-2-thiouridine modification at tRNA U34
		pth	*mge*	peptidyl-tRNA hydrolase (EC:3.1.1.29)
		$rnpA$	*mge*	ribonuclease P protein component (EC:3.1.26.5)
Transport (23 genes)	$aroP$	$aroP$	*eco*	aromatic amino acid transporter
	$bgtT$	$bgtA$	*syc*	ABC-type permease for basic amino acids and glutamine

Table B.3 (Continued)

Category	MCM ID	Gene(s)	Species	Function
		bgtB	*syc*	ABC-type permease for basic amino acids and glutamine
	bztD	*bztD*	*rsp*	ABC glutamate/glutamine/aspartate/asparagine transporter, ATPase subunit BztD
	kup	*kup*	*eco*	potassium transporter
	lctP	*lctP*	*bsu*	L-lactate permease
	livF	*livF*	*eco*	leucine/isoleucine/valine transporter subunit
	metT	*metI*	*eco*	DL-methionine transporter subunit
		metN	*eco*	DL-methionine transporter subunit
		metQ	*eco*	DL-methionine transporter subunit
	mgtA	*mgtA*	*eco*	magnesium transporter (EC:3.6.3.1)
	mntH	*mntH*	*eco*	manganese/divalent cation transporter
	natT	*natA*	*bsu*	Na+ ABC efflux transporter (ATP-binding protein)
		natB	*bsu*	Na+ ABC efflux transporter (permease)
		natC	*syc*	ABC-type Nat permease for neutral amino acids

Table B.3 (Continued)

Category	MCM ID	Gene(s)	Species	Function
		$natD$	*syc*	integral membrane protein of the ABC-type Nat permease for neutral amino acids NatD
	$nhaB$	$nhaB$	*eco*	sodium:proton antiporter
	$pitA$	$pitA$	*eco*	phosphate transporter, low-affinity
	$ptsT$	$ptsG$	*mge*	PTS system, glucose-specific IIABC component (EC:2.7.1.69)
		$ptsH$	*mge*	phosphocarrier protein HPr
		$ptsI$	*mge*	phosphoenolpyruvate-protein phosphotransferase (EC:2.7.3.9)
	$sstT$	$sstT$	*eco*	sodium:serine/threonine symporter
	$tcyP$	$tcyP$	*bsu*	sodium-cystine symporter

REFERENCES

Gil, R., Silva, F. J., Pereto, J., and Moya, A. (2004). Determination of the core of a minimal bacterial gene set. *Microbiology and Molecular Biology Reviews*, 68(3), 518–537.

Kanehisa, M. and Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1), 27–30.

APPENDIX C

METABOLIC PATHWAYS IN THE MINIMAL CELL MODEL

The Minimal Cell Model (MCM) contains detailed descriptions of glycolysis, the pentose phosphate pathway, lipid biosynthesis, nucleotide biosynthesis, cofactor metabolism, and energy metabolism via fermentation. The Reaction module is dedicated to defining the reactions of metabolism. The overall metabolism of the MCM is presented in Section 4.14 and summarized in Figure 4.2. Each of the submodules of metabolism are illustrated in this Appendix in Figures C.1-C.8. The central metabolic pathways included in the MCM are based on the minimal gene set proposed by Gil et al. (2004).

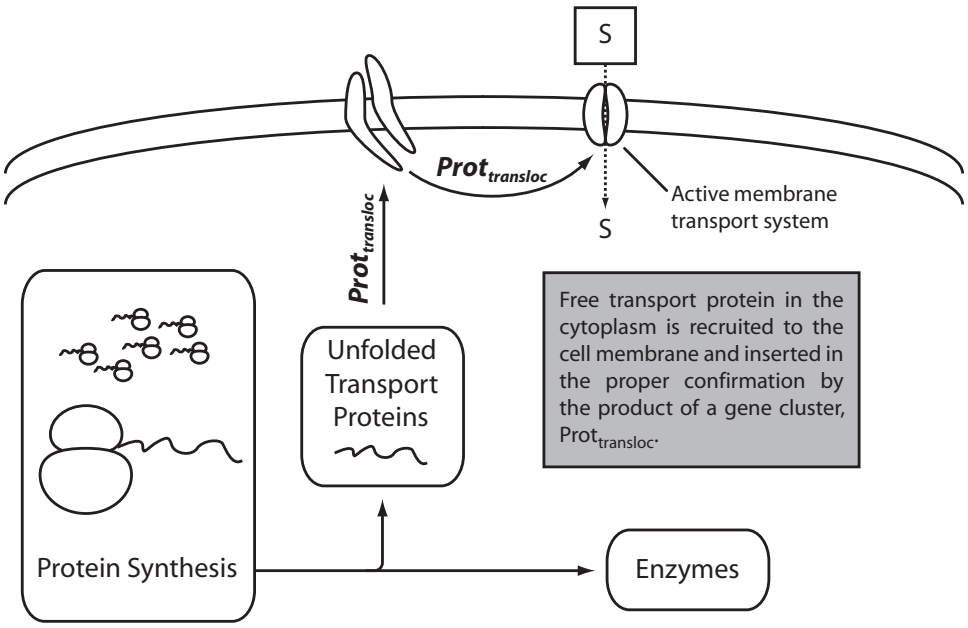


Figure C.1: Transporter assembly in the Minimal Cell Model. The coarse-grained model for membrane protein insertion in the MCM. Gene cluster $prot_{\mathrm{transloc}}$ includes the genes *ffh*, *ftsY*, *secA*, *secE*, and *secY*.

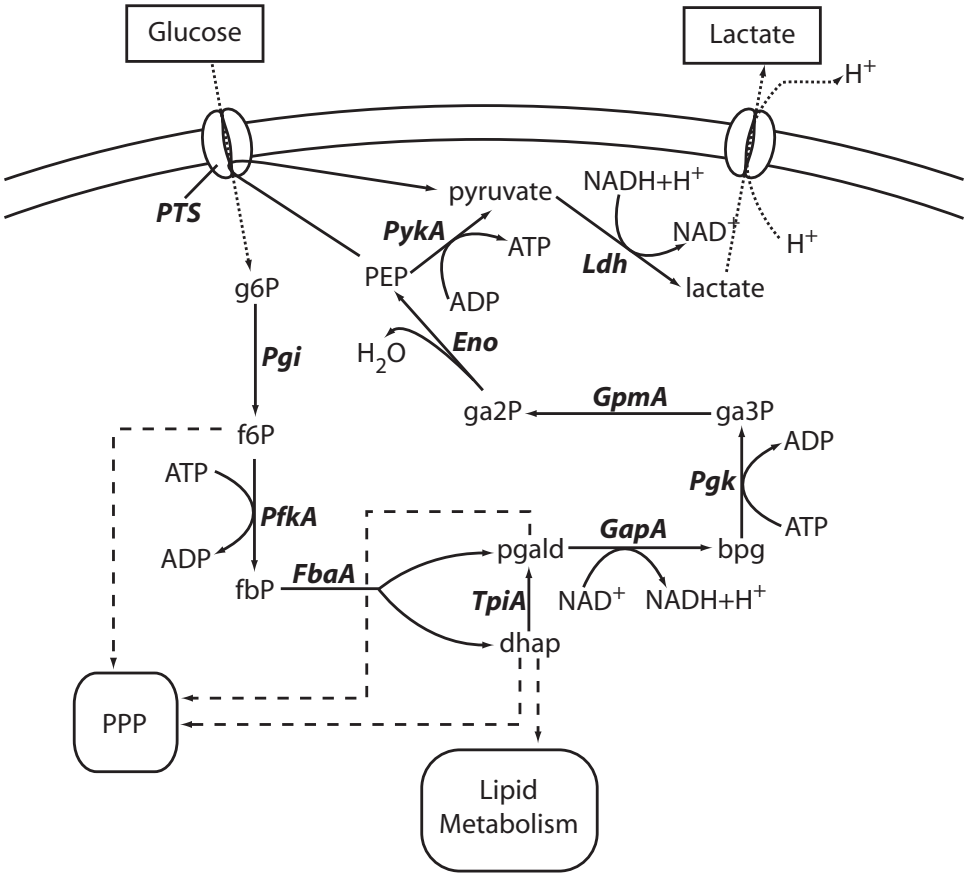


Figure C.2: Glycolysis reactions included in the Minimal Cell Model. Solid arrows represent mass flow, while dashed arrows represent connections to other metabolic pathways. Labels in *italic* are enzymes, defined as follows: Pgi, glucose-6-phosphate isomerase; Pkfa, 6-phosphfructokinase; FbaA, fructose-1,6-bisphosphate aldolase; TpiA, triose phosphate isomerase, GapA, glyceraldehyde 3-phosphate dehydrogenase; Pgk, phosphoglycerate kinase; GpmA, phosphglycerate mutase; Eno, enolase; PykA, pyruvate kinase; Ldh, lactate dehydrogenase; PTS, phosphotransferase system. Chemical species abbreviations defined in Table E.1.

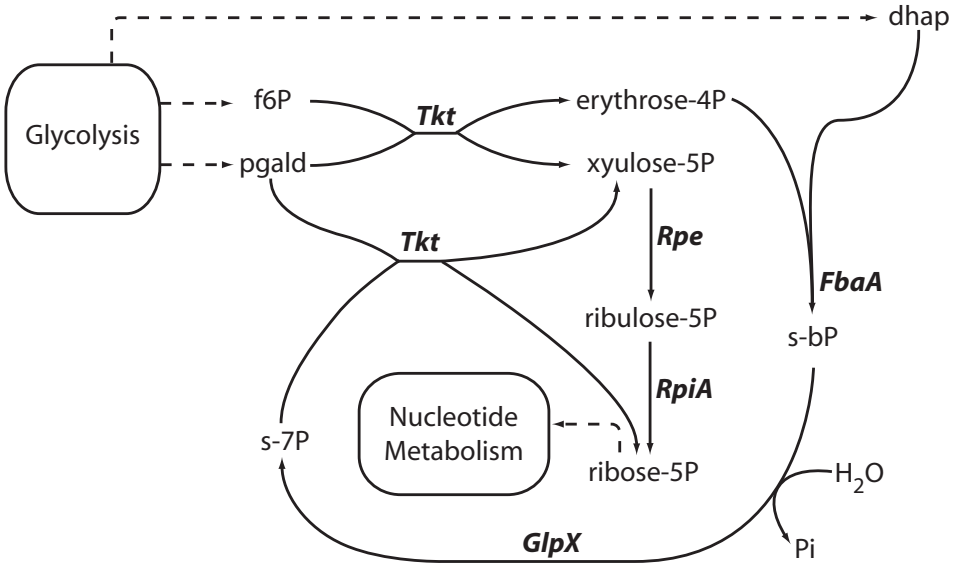


Figure C.3: Pentose phosphate pathway (PPP) reactions included in the Minimal Cell Model. Solid arrows represent mass flow, while dashed arrows represent connections to other metabolic pathways. Labels in *italic* are enzymes, defined as follows: Tkt, transketolase; FbaA, fructose-1,6-bisphosphate aldolase; GlpX, sedoheptulose-bisphosphatase; Rpe, ribulose-phosphate 3-epimerase; RpiA, ribose 5-phosphate isomerase. Nonstandard chemical abbreviations are: s-bP sedoheptulose 1,7-bisphosphate; sedoheptulose 7-phosphate. The remaining chemical species abbreviations are defined in Table E.1.

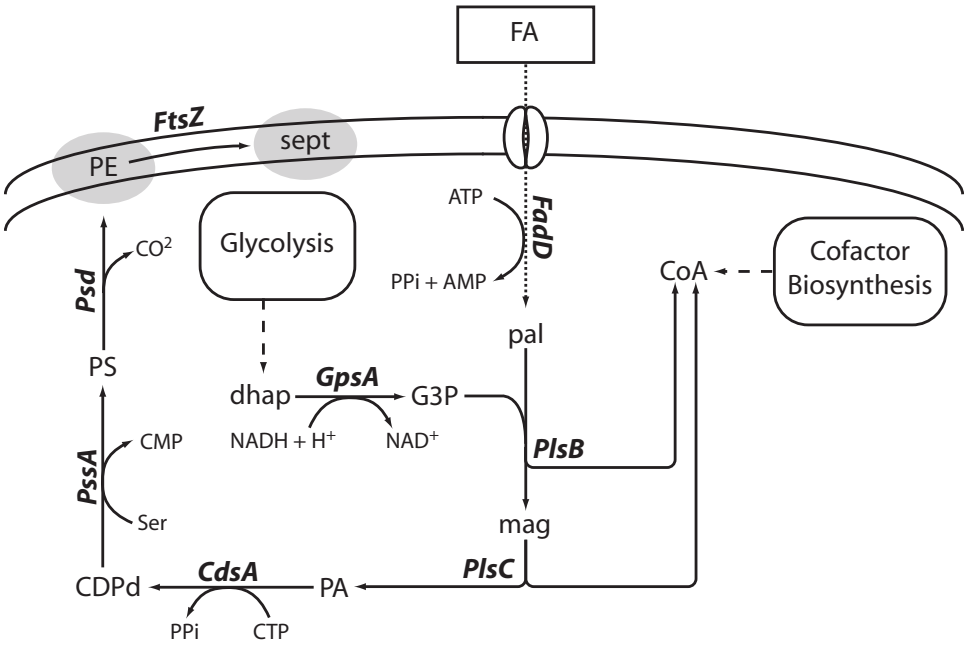


Figure C.4: Lipid biosynthesis reactions included in the Minimal Cell Model. Solid arrows represent mass flow, while dashed arrows represent connections to other metabolic pathways or transport processes. The FtsZ reaction, which recruits lipid membranes PE to the septum at the midcell region, is not active until chromosome replication terminates. Labels in *italic* are enzymes, defined as follows: FadD, acyl-CoA synthase; PlsB, *sn*-glycerol-3-phosphate acyltransferase; PlsC, 1-acyl-*sn*-glycerol-3-phosphate acyltransferase; GpsA, *sn*-glycerol-3-phosphate dehydrogenase; CdsA, phosphatidate cytidyltransferase; PssA, phosphatidylserine synthase; Psd phosphatidylserine synthase; FtsZ, cytoskeletal cell division protein. Nonstandard chemical abbreviations are: FA, external palmitate; pal, palmitoyl CoA; mag, lysophosphatidate; PA, phosphatidate; CDPd, CDP-diacylglycerol; PS, phosphatidylserine; PE phosphatidylethanolamine. The remaining chemical species abbreviations are defined in Table E.1.

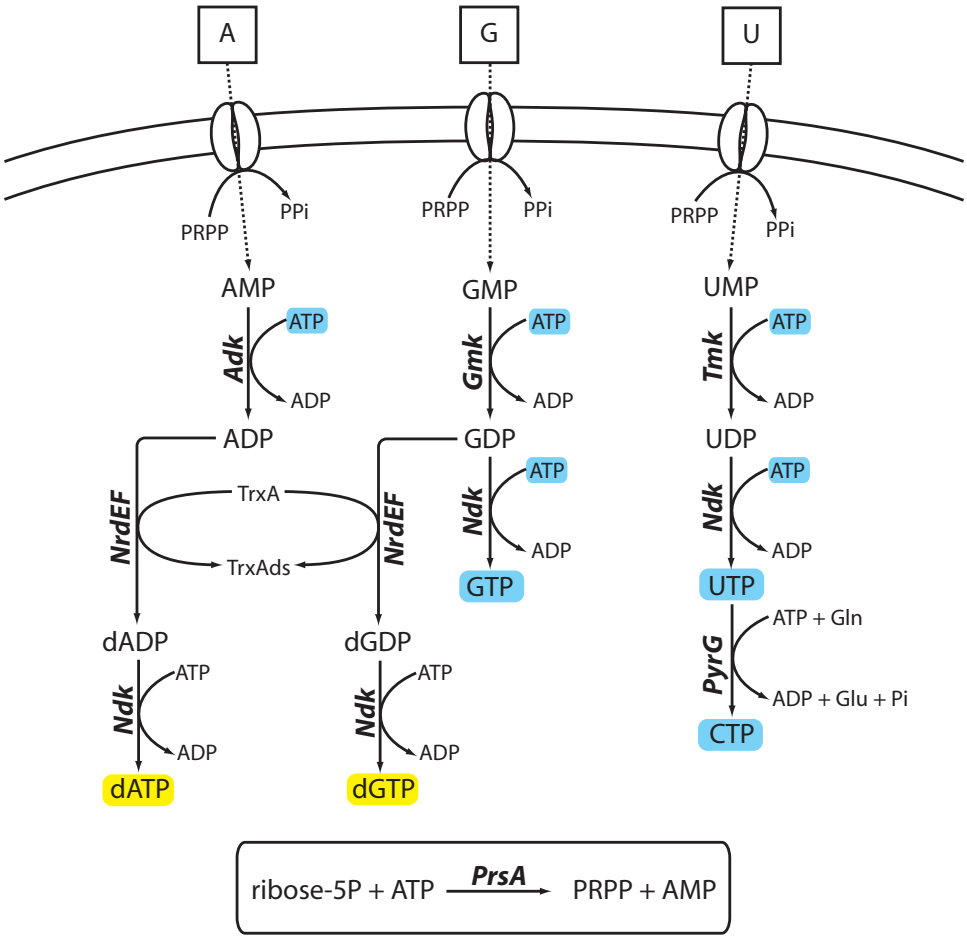


Figure C.5: Ribonucleotide biosynthesis reactions included in the Minimal Cell Model. Solid arrows represent mass flow, while dashed arrows represent connections to other metabolic pathways or transport processes. Blue labels refer to ribonucleotide triphosphates, while yellow labels refer to deoxyribonucleotide triphosphates. Labels in *italic* are enzymes, defined as follows: Adk, adenylate kinase; NrdEF, ribonucleoside-diphosphate reductase; Ndk, nucleoside diphosphate kinase; Gmk, guanylate kinase; Tmk, thymidylate kinase; PyrG, CTP synthase; PrsA, phosphribosylpyrophosphate synthase. Chemical species abbreviations are defined in Table E.1.

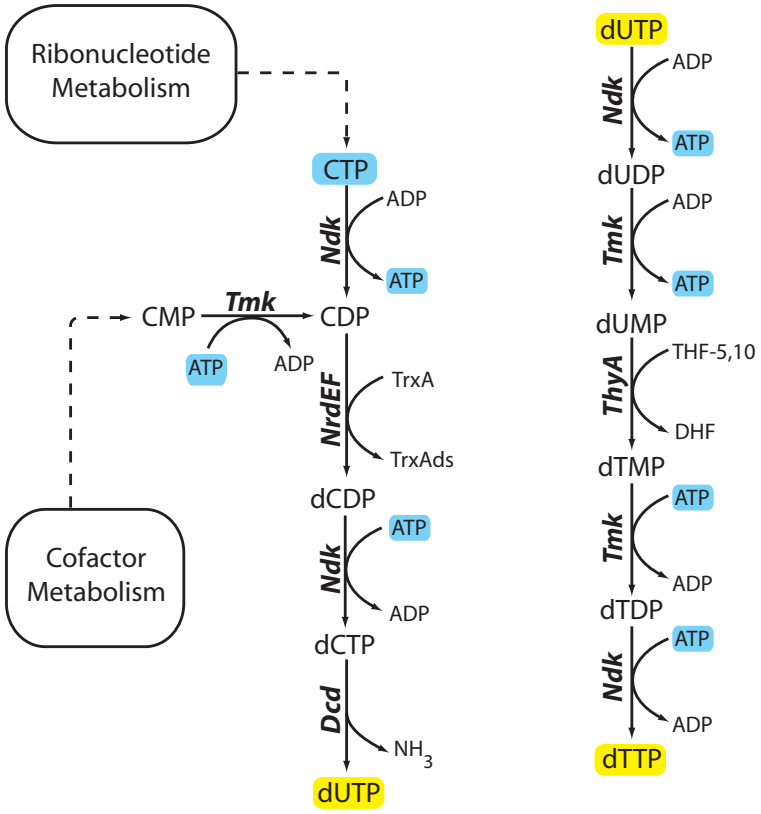


Figure C.6: Deoxyribonucleotide biosynthesis reactions included in the Minimal Cell Model. Solid arrows represent mass flow, while dashed arrows represent connections to other metabolic pathways or transport processes. Blue labels refer to ribonucleotide triphosphates, while yellow labels refer to deoxyribonucleotide triphosphates. Labels in *italic* are enzymes, defined as follows: Ndk, nucleoside diphosphate kinase; Tmk, thymidylate kinase; NrdEF, ribonucleoside-diphosphate reductase; Dcd, dCTP deaminase; ThyA thymidylate synthase. Chemical species abbreviations are defined in Table E.1.

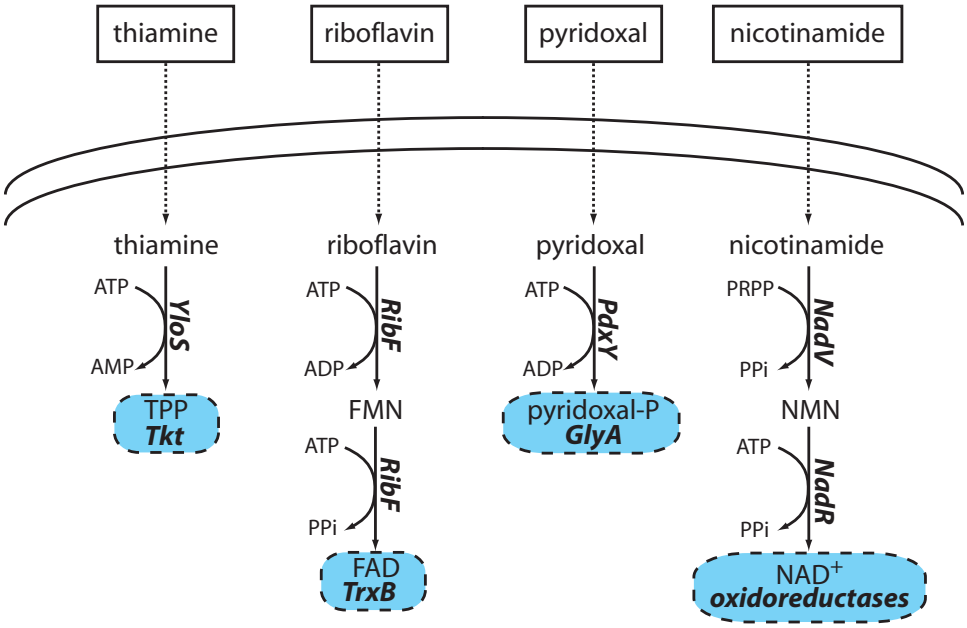


Figure C.7: Cofactor biosynthesis reaction pathways included in the Minimal Cell Model (1 of 2). Solid arrows represent mass flow, while dashed arrows represent connections to other metabolic pathways or transport processes. Blue shading indicates a relationship between a cofactor and an enzyme or group of enzymes. Labels in *italic* are enzymes, defined as follows: YloS, thiamine pyrophosphokinase; RibF, riboflavin kinase, FMN adenyltransferase; PdxY, pyridoxal kinase; NadV, nicotinamide phosphoribosyltransferase; NadR, adenilyl transferase. Chemical species abbreviations are defined in Table E.1.

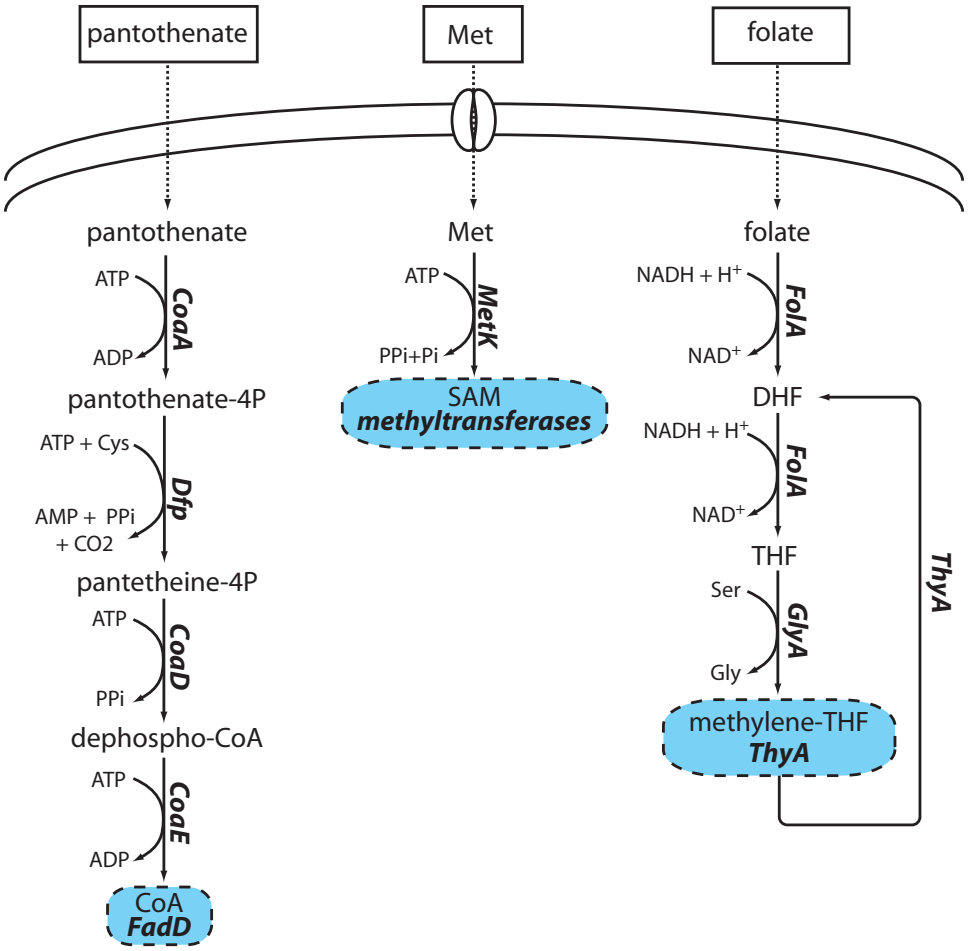


Figure C.8: Cofactor biosynthesis reaction pathways included in the Minimal Cell Model (2 of 2). Solid arrows represent mass flow, while dashed arrows represent connections to other metabolic pathways or transport processes. Blue shading indicates a relationship between a cofactor and an enzyme or group of enzymes. Note that the dUMP reactant and the dTMP product of the ThyA reaction are not pictured. Labels in *italic* are enzymes, defined as follows: CoaA, pantothenate kinase; Dfp, phosphopantothenate cysteine ligase, 4' phospho-pantothenyl-L-cysteine decarboxylase; CoaD, 4'-phospho-pantetheine adenyltransferase; CoaE, dephosphocoenzyme A kinase; MetK, methionine adenyltransferase; FolA, dihydrofolate reductase; GlyA, glycine hydroxymethyltransferase; ThyA, thymidylate synthase. Chemical species abbreviations are defined in Table E.1.

REFERENCES

Gil, R., Silva, F. J., Pereto, J., and Moya, A. (2004). Determination of the core of a minimal bacterial gene set. *Microbiology and Molecular Biology Reviews*, 68(3), 518–537.

APPENDIX D

MINIMAL CELL EXTERNAL ENVIRONMENT

A chemically and genomically detailed model of a minimal cell growing in an optimally supportive culture environment has been created. The Minimal Cell Model (MCM) is defined to exist in a constant, benign environment with optimal concentrations of all its required nutrients, pH, temperature, and dilution of any waste products. The cell concentration in this environment is considered to be low enough that the nutrients are never significantly diluted. Alternatively the cell could be considered to be growing in a continuous flow stirred tank reactor (CFSTR) that is operating at steady-state.

The 38 compounds present in the medium are listed in Tables D.1 and D.2. Concentrations proposed for defined media for *Mycoplasma* strain Y (which is similar to *M. mycoides*) for glucose; free bases A, G, and U; some cofactor precursors; and the amino acids were used as the basis for the MCM's external medium (Rodwell, 1969).

No suitable reference for the concentration of folic acid, fatty acids, pantothenic acid, or inorganic ions was available, so their initial external concentrations were set to $1 \times 10^{-3} \frac{\text{gm}}{\text{mL}}$. Because the external environment is assumed to be constant, changes in the concentrations of external nutrients could be compensated for by changes in the rate constants for transport reactions.

Table D.1: Extracellular amino acids in the medium compartment. ID is the string identifier used for each chemical within the model. External concentrations are set high enough so that the medium is considered to be optimally supportive. Many of the initial concentrations were based on a defined media for *Mycoplasma* Strain Y (Rodwell, 1969). For Asp, Tyr, and Gln concentrations of 1.0 mM (converted into mass based concentrations here) have been assumed.

Name	**ID**	**Concentration ($\frac{\mathbf{gm}}{\mathbf{mL}}$)**
alanine*	$\mathrm{Ala}_{\mathrm{ext}}$	1.8×10^{-4}
arginine*	$\mathrm{Arg}_{\mathrm{ext}}$	1.7×10^{-4}
asparagine*	$\mathrm{Asn}_{\mathrm{ext}}$	1.5×10^{-4}
aspartate*	$\mathrm{Asp}_{\mathrm{ext}}$	1.3×10^{-4}
cystine*	$\mathrm{Cys}_{\mathrm{ext}}$	2.4×10^{-4}
glutamicacid*	$\mathrm{Glu}_{\mathrm{ext}}$	1.5×10^{-4}
glutamine*	$\mathrm{Gln}_{\mathrm{ext}}$	1.5×10^{-4}
glycine*	$\mathrm{Gly}_{\mathrm{ext}}$	1.5×10^{-4}
histidine*	$\mathrm{His}_{\mathrm{ext}}$	1.6×10^{-4}
isoleucine*	$\mathrm{Ile}_{\mathrm{ext}}$	1.3×10^{-4}
leucine*	$\mathrm{Leu}_{\mathrm{ext}}$	1.3×10^{-4}
lysine*	$\mathrm{Lys}_{\mathrm{ext}}$	1.5×10^{-4}
methionine*	$\mathrm{Met}_{\mathrm{ext}}$	3.0×10^{-4}
phenylalanine*	$\mathrm{Phe}_{\mathrm{ext}}$	3.3×10^{-4}
proline*	$\mathrm{Pro}_{\mathrm{ext}}$	1.2×10^{-4}
serine*	$\mathrm{Ser}_{\mathrm{ext}}$	2.1×10^{-4}
threonine*	$\mathrm{Thr}_{\mathrm{ext}}$	2.4×10^{-4}
tryptophan*	$\mathrm{Trp}_{\mathrm{ext}}$	4.1×10^{-4}
tyrosine*	$\mathrm{Tyr}_{\mathrm{ext}}$	1.8×10^{-4}
valine*	$\mathrm{Val}_{\mathrm{ext}}$	2.3×10^{-4}

Table D.2: Extracellular species present in the medium, aside from amino acids. See Table D.1 for amino acid concentrations. Species ID is the string identifier used within the model. External concentrations are set high enough so that the medium is considered to be optimally supportive. Many of the initial concentrations were based on a defined media for *Mycoplasma* Strain Y (Rodwell, 1969). For inorganic ions and some precursors of cofactor biosynthesis concentrations of 1.0 $\frac{\text{gm}}{\text{mL}}$ have been assumed.

Species Name	**Species ID**	**Concentration ($\frac{\text{gm}}{\text{mL}}$)**
K*	K_{ext}	1.0×10^{-3}
Mg*	Mg_{ext}	1.0×10^{-3}
Mn*	Mn_{ext}	1.0×10^{-3}
Na*	Na_{ext}	1.0×10^{-3}
Pi*	Pi_{ext}	1.4×10^{-2}
adenine*	A_{ext}	1.0×10^{-5}
fattyacids*	FA_{ext}	1.0×10^{-3}
folate∗	$\text{folate}_{\text{ext}}$	1.0×10^{-3}
glucose*	A2_{ext}	7.0×10^{-3}
guanine*	G_{ext}	1.0×10^{-5}
hydrogen*	H_{ext}	1.0×10^{-3}
lactate*	$\text{lactate}_{\text{ext}}$	1.0×10^{-4}
nicotinamide*	$\text{nicotinamide}_{\text{ext}}$	1.0×10^{-6}
pantothenate*	$\text{pantothenate}_{\text{ext}}$	1.0×10^{-3}
pyridoxal*	$\text{pyridoxal}_{\text{ext}}$	1.0×10^{-3}
riboflavin*	$\text{riboflavin}_{\text{ext}}$	1.0×10^{-6}
thiamine*	$\text{thiamine}_{\text{ext}}$	1.0×10^{-6}
uracil*	U_{ext}	1.0×10^{-5}

REFERENCES

Rodwell, A. W. (1969). A defined medium for *Mycoplasma* strain Y. *Journal of General Microbiology*, 58, 39–&.

APPENDIX E

INITIAL CONDITIONS FOR THE MINIMAL CELL MODEL

A chemically detailed model of a bacterial cell must have the initial mass of all its chemical species specified. For many chemical species, even average cell cycle values are not known, let alone detailed concentration information as a function of the cell cycle progression. To obtain initial conditions for the Minimal Cell Model (MCM), we make use of data for groups of chemical species published for *E. coli* and make assumptions about how these groups are subdivided in our hypothetical cell (Neidhardt, 1996). Because there is no experimental analog for a minimal cell, we propose that using composition data measured in *E. coli* is a valid first-approximation because it will have a similar chemical make-up to other chemoheterotrophic bacteria. The procedure used to calculate the initial conditions is presented in full in Section 4.5.1. Table E.1 shows the results of applying that procedure in the MCM.

Table E.1: Chemical species and their initial masses in the Minimal Cell Model. Id refers to the variable name for the chemical species in the MCM, while Name is a more descriptive identifier for the chemical. Type is a simple description of what kind of molecule the species represents. Location is either Cytoplasm or Membrane. The initial mass is given in pg and determined as described in the text. Note that for certain chemical species involved in Demand objects (Section 4.18), the initial condition is shifted by a small amount (<1%) to ensure that one of the chemicals is initially limiting.

Id	Name	Type	Location	Initial Mass (pg)
adk_mRNA	adk_{mRNA}	mRNA	Cytoplasm	5.58×10^{-6}
Adk	Adk	Protein	Cytoplasm	3.72×10^{-4}
ADP	adenosine diphosphate	Nucleotide	Cytoplasm	6.38×10^{-5}
Ala_tRNA	Ala_{tRNA}	Amino-Acyl tRNA	Cytoplasm	1.20×10^{-4}
alaS_mRNA	$alaS_{mRNA}$	mRNA	Cytoplasm	5.58×10^{-6}
AlaS	AlaS	Protein	Cytoplasm	3.72×10^{-4}
Ala	alanine	Amino Acid	Cytoplasm	6.38×10^{-5}
AMP	adenosine monophosphate	Nucleotide	Cytoplasm	6.38×10^{-5}
Arg_tRNA	Arg_{tRNA}	Amino-Acyl tRNA	Cytoplasm	1.25×10^{-4}
argS_mRNA	$argS_{mRNA}$	mRNA	Cytoplasm	5.58×10^{-6}
ArgS	ArgS	Protein	Cytoplasm	3.72×10^{-4}
Arg	arginine	Amino Acid	Cytoplasm	6.38×10^{-5}
aroP_mRNA	$aroP_{mRNA}$	mRNA	Cytoplasm	5.58×10^{-6}
AroP	AroP	Protein	Cytoplasm	3.72×10^{-4}
Asn_tRNA	Asn_{tRNA}	Amino-Acyl tRNA	Cytoplasm	1.31×10^{-4}

Table E.1 (Continued)

Id	Name	Type	Location	Initial Mass (pg)
asnS_mRNA	$asnS_{mRNA}$	mRNA	Cytoplasm	5.58×10^{-6}
AsnS	AsnS	Protein	Cytoplasm	3.72×10^{-4}
Asn	asparagine	Amino Acid	Cytoplasm	6.38×10^{-5}
Asp_tRNA	Asp_{tRNA}	Amino-Acyl tRNA	Cytoplasm	1.36×10^{-4}
aspS_mRNA	$aspS_{mRNA}$	mRNA	Cytoplasm	5.58×10^{-6}
AspS	AspS	Protein	Cytoplasm	3.72×10^{-4}
Asp	aspartate	Amino Acid	Cytoplasm	6.38×10^{-5}
ATP	adenosine triphosphate	Nucleotide	Cytoplasm	6.57×10^{-5}
bgtT_mRNA	$bgtT_{mRNA}$	mRNA	Cytoplasm	1.12×10^{-5}
BgtT	BgtT	Protein	Cytoplasm	7.43×10^{-4}
bpg	1,3-bisphosphoglycerate	Glycolytic	Cytoplasm	6.38×10^{-5}
bztD_mRNA	$bztD_{mRNA}$	mRNA	Cytoplasm	5.58×10^{-6}
BztD	BztD	Protein	Cytoplasm	3.72×10^{-4}
CDPd	CDP-diacylglycerol	Lipid	Cytoplasm	6.38×10^{-5}
CDP	cytidine diphosphate	Nucleotide	Cytoplasm	6.38×10^{-5}
cdsA_mRNA	$cdsA_{mRNA}$	mRNA	Cytoplasm	5.58×10^{-6}
CdsA	CdsA	Protein	Cytoplasm	3.72×10^{-4}
CMP	cytidine monophosphate	Nucleotide	Cytoplasm	6.38×10^{-5}

Table E.1 (Continued)

Id	Name	Type	Location	Initial Mass (pg)
coaA_mRNA	$coaA_{mRNA}$	mRNA	Cytoplasm	5.58×10^{-6}
CoaA	CoaA	Protein	Cytoplasm	3.72×10^{-4}
coaD_mRNA	$coaD_{mRNA}$	mRNA	Cytoplasm	5.58×10^{-6}
CoAdp	dephospho-CoA	Cofactor	Cytoplasm	6.38×10^{-5}
CoaD	CoaD	Protein	Cytoplasm	3.72×10^{-4}
coaE_mRNA	$coaE_{mRNA}$	mRNA	Cytoplasm	5.58×10^{-6}
CoaE	CoaE	Protein	Cytoplasm	3.72×10^{-4}
CoA	conenzyme A	Cofactor	Cytoplasm	6.38×10^{-5}
CTP	cytidine triphosphate	Nucleotide	Cytoplasm	6.51×10^{-5}
Cys_tRNA	Cys_{tRNA}	Amino-Acyl tRNA	Cytoplasm	1.32×10^{-4}
cysS_mRNA	$cysS_{mRNA}$	mRNA	Cytoplasm	5.58×10^{-6}
CysS	CysS	Protein	Cytoplasm	3.72×10^{-4}
Cys	cystine	Amino Acid	Cytoplasm	6.38×10^{-5}
dADP	deoxyadenosine diphosphate	Nucleotide	Cytoplasm	6.38×10^{-5}
dATP	deoxyadenosine triphosphate	Nucleotide	Cytoplasm	6.51×10^{-5}
dcd_mRNA	dcd_{mRNA}	mRNA	Cytoplasm	5.58×10^{-6}
dCDP	deoxycytidine diphosphate	Nucleotide	Cytoplasm	6.38×10^{-5}
Dcd	Dcd	Protein	Cytoplasm	3.72×10^{-4}

Table E.1 (Continued)

Id	Name	Type	Location	Initial Mass (pg)
dCTP	deoxycytidine triphosphate	Nucleotide	Cytoplasm	6.57×10^{-5}
deg_M1_mRNA	$\text{deg}_{\text{M1-mRNA}}$	mRNA	Cytoplasm	1.67×10^{-5}
Deg_M1	Deg_{M1}	Protein	Cytoplasm	1.12×10^{-3}
deg_RNA_mRNA	$\text{deg}_{\text{RNA-mRNA}}$	mRNA	Cytoplasm	1.12×10^{-5}
Deg_RNA	Deg_{RNA}	Protein	Cytoplasm	7.43×10^{-4}
dfp_mRNA	dfp_{mRNA}	mRNA	Cytoplasm	5.58×10^{-6}
Dfp	Dfp	Protein	Cytoplasm	3.72×10^{-4}
dGDP	deoxyguanosine diphosphate	Nucleotide	Cytoplasm	6.38×10^{-5}
dGTP	deoxyguanosine triphosphate	Nucleotide	Cytoplasm	6.44×10^{-5}
dhap	dihydroxyacetone phosphate	Glycolytic	Cytoplasm	6.38×10^{-5}
DHF	dihydrofolate	Cofactor	Cytoplasm	6.38×10^{-5}
dna_rep_mRNA	$\text{dna}_{\text{rep-mRNA}}$	mRNA	Cytoplasm	1.67×10^{-5}
Dna_rep	Dna_{rep}	Protein	Cytoplasm	1.12×10^{-3}
DnaB_boundto_Ori	$\text{DnaB}_{\text{boundto-Ori}}$	Protein	Cytoplasm	2.40×10^{-6}
dnaB_mRNA	$\text{dnaB}_{\text{mRNA}}$	mRNA	Cytoplasm	5.58×10^{-6}
DnaB	DnaB	Protein	Cytoplasm	3.72×10^{-4}
DnaG_boundto_Ori	$\text{DnaG}_{\text{boundto-Ori}}$	Protein	Cytoplasm	2.40×10^{-6}
dnaG_mRNA	$\text{dnaG}_{\text{mRNA}}$	mRNA	Cytoplasm	5.58×10^{-6}

Table E.1 (Continued)

Id	Name	Type	Location	Initial Mass (pg)
DnaG	DnaG	Protein	Cytoplasm	3.72×10^{-4}
dTDP	deoxythymidine diphosphate	Nucleotide	Cytoplasm	6.38×10^{-5}
dTMP	deoxythymidine monophosphate	Nucleotide	Cytoplasm	6.38×10^{-5}
dTTP	deoxythymidine triphosphate	Nucleotide	Cytoplasm	6.38×10^{-5}
dUDP	deoxyuridine diphosphate	Nucleotide	Cytoplasm	6.38×10^{-5}
dUMP	deoxyuridine monophosphate	Nucleotide	Cytoplasm	6.38×10^{-5}
dUTP	deoxyuridine triiphosphate	Nucleotide	Cytoplasm	6.38×10^{-5}
eno_mRNA	eno_{mRNA}	mRNA	Cytoplasm	5.58×10^{-6}
Eno	Eno	Protein	Cytoplasm	3.72×10^{-4}
erythrose4P	erythrose-4-P	Pentose Phosphate	Cytoplasm	6.38×10^{-5}
f6P	fructose 6phosphate	Glycolytic	Cytoplasm	6.38×10^{-5}
fadD_mRNA	$\text{fadD}_{\text{mRNA}}$	mRNA	Cytoplasm	5.58×10^{-6}
FadD	FadD	Protein	Cytoplasm	3.72×10^{-4}
FAD	flavin adenine dinucleotide	Cofactor	Cytoplasm	6.38×10^{-5}
fbaA_mRNA	$\text{fbaA}_{\text{mRNA}}$	mRNA	Cytoplasm	5.58×10^{-6}
FbaA	FbaA	Protein	Cytoplasm	3.72×10^{-4}
fbP	fructose-1-6-P	Glycolytic	Cytoplasm	6.38×10^{-5}
FMN	flavin mononucleotide	Cofactor	Cytoplasm	6.38×10^{-5}

Table E.1 (Continued)

Id	Name	Type	Location	Initial Mass (pg)
folA_mRNA	$folA_{mRNA}$	mRNA	Cytoplasm	5.58×10^{-6}
folate	folate	Cofactor	Cytoplasm	1.02×10^{-4}
FolA	FolA	Protein	Cytoplasm	3.72×10^{-4}
ftsZ_mRNA	$ftsZ_{mRNA}$	mRNA	Cytoplasm	5.58×10^{-6}
FtsZ	FtsZ	Protein	Cytoplasm	3.72×10^{-4}
G3P	glycerol-3P	Lipid	Cytoplasm	6.38×10^{-5}
g6P	glucose-6-P	Glycolytic	Cytoplasm	6.38×10^{-5}
ga2P	glyceraldehyde-2P	Glycolytic	Cytoplasm	6.38×10^{-5}
ga3P	glyceraldehyde-3P	Glycolytic	Cytoplasm	6.38×10^{-5}
gapA_mRNA	$gapA_{mRNA}$	mRNA	Cytoplasm	5.58×10^{-6}
GapA	GapA	Protein	Cytoplasm	3.72×10^{-4}
GDP	guanosine diphosphate	Nucleotide	Cytoplasm	6.38×10^{-5}
Gln_tRNA	Gln_{tRNA}	Amino-Acyl tRNA	Cytoplasm	1.35×10^{-4}
glnS_mRNA	$glnS_{mRNA}$	mRNA	Cytoplasm	5.58×10^{-6}
GlnS	GlnS	Protein	Cytoplasm	3.72×10^{-4}
Gln	glutamine	Amino Acid	Cytoplasm	6.38×10^{-5}
glpX_mRNA	$glpX_{mRNA}$	mRNA	Cytoplasm	5.58×10^{-6}
GlpX	GlpX	Protein	Cytoplasm	3.72×10^{-4}

Table E.1 (Continued)

Id	Name	Type	Location	Initial Mass (pg)
gltX_mRNA	$gltX_{mRNA}$	mRNA	Cytoplasm	5.58×10^{-6}
GltX	GltX	Protein	Cytoplasm	3.72×10^{-4}
Glu_tRNA	Glu_{tRNA}	Amino-Acyl tRNA	Cytoplasm	1.39×10^{-4}
Glu	glutamic acid	Amino Acid	Cytoplasm	6.38×10^{-5}
Gly_tRNA	Gly_{tRNA}	Amino-Acyl tRNA	Cytoplasm	1.28×10^{-4}
glyA_mRNA	$glyA_{mRNA}$	mRNA	Cytoplasm	5.58×10^{-6}
GlyA	GlyA	Protein	Cytoplasm	3.72×10^{-4}
glyS_mRNA	$glyS_{mRNA}$	mRNA	Cytoplasm	5.58×10^{-6}
GlyS	GlyS	Protein	Cytoplasm	3.72×10^{-4}
Gly	glycine	Amino Acid	Cytoplasm	6.38×10^{-5}
gmk_mRNA	gmk_{mRNA}	mRNA	Cytoplasm	5.58×10^{-6}
Gmk	Gmk	Protein	Cytoplasm	3.72×10^{-4}
GMP	guanosine monophosphate	Nucleotide	Cytoplasm	6.38×10^{-5}
gpmA_mRNA	$gpmA_{mRNA}$	mRNA	Cytoplasm	5.58×10^{-6}
GpmA	GpmA	Protein	Cytoplasm	3.72×10^{-4}
gpsA_mRNA	$gpsA_{mRNA}$	mRNA	Cytoplasm	5.58×10^{-6}
GpsA	GpsA	Protein	Cytoplasm	3.72×10^{-4}
GTP	guanosine triphosphate	Nucleotide	Cytoplasm	6.38×10^{-5}

Table E.1 (Continued)

Id	Name	Type	Location	Initial Mass (pg)
His_tRNA	His_{tRNA}	Amino-Acyl tRNA	Cytoplasm	1.30×10^{-4}
hisS_mRNA	$\text{hisS}_{\text{mRNA}}$	mRNA	Cytoplasm	5.58×10^{-6}
HisS	HisS	Protein	Cytoplasm	3.72×10^{-4}
His	histidine	Amino Acid	Cytoplasm	6.38×10^{-5}
hpt_mRNA	hpt_{mRNA}	mRNA	Cytoplasm	5.58×10^{-6}
Hpt	Hpt	Protein	Cytoplasm	3.72×10^{-4}
HupA_boundto_Ori	$\text{HupA}_{\text{boundto}-\text{Ori}}$	Protein	Cytoplasm	2.40×10^{-6}
hupA_mRNA	$\text{hupA}_{\text{mRNA}}$	mRNA	Cytoplasm	5.58×10^{-6}
HupA	HupA	Protein	Cytoplasm	3.72×10^{-4}
Ile_tRNA	Ile_{tRNA}	Amino-Acyl tRNA	Cytoplasm	1.23×10^{-4}
ileS_mRNA	$\text{ileS}_{\text{mRNA}}$	mRNA	Cytoplasm	5.58×10^{-6}
IleS	IleS	Protein	Cytoplasm	3.72×10^{-4}
Ile	isoleucine	Amino Acid	Cytoplasm	6.38×10^{-5}
kup_mRNA	kup_{mRNA}	mRNA	Cytoplasm	5.58×10^{-6}
Kup	Kup	Protein	Cytoplasm	3.72×10^{-4}
lactate	lactate	Glycolytic	Cytoplasm	6.38×10^{-5}
lctP_mRNA	$\text{lctP}_{\text{mRNA}}$	mRNA	Cytoplasm	5.58×10^{-6}
LctP	LctP	Protein	Cytoplasm	3.72×10^{-4}

Table E.1 (Continued)

Id	Name	Type	Location	Initial Mass (pg)
ldh_mRNA	ldh_{mRNA}	mRNA	Cytoplasm	5.58×10^{-6}
Ldh	Ldh	Protein	Cytoplasm	3.72×10^{-4}
Leu_tRNA	Leu_{tRNA}	Amino-Acyl tRNA	Cytoplasm	1.33×10^{-4}
leuS_mRNA	$leuS_{mRNA}$	mRNA	Cytoplasm	5.58×10^{-6}
LeuS	LeuS	Protein	Cytoplasm	3.72×10^{-4}
Leu	leucine	Amino Acid	Cytoplasm	6.38×10^{-5}
livF_mRNA	$livF_{mRNA}$	mRNA	Cytoplasm	5.58×10^{-6}
LivF	LivF	Protein	Cytoplasm	3.72×10^{-4}
Lys_tRNA	Lys_{tRNA}	Amino-Acyl tRNA	Cytoplasm	1.38×10^{-4}
lysS_mRNA	$lysS_{mRNA}$	mRNA	Cytoplasm	5.58×10^{-6}
LysS	LysS	Protein	Cytoplasm	3.72×10^{-4}
Lys	lysine	Amino Acid	Cytoplasm	6.38×10^{-5}
M2_RTM	$RNA_{stable,mature}$	rRNA	Cytoplasm	1.67×10^{-5}
M3	DNA	DNA	Cytoplasm	3.77×10^{-4}
mag	mag	Lipid	Cytoplasm	6.38×10^{-5}
map_mRNA	map_{mRNA}	mRNA	Cytoplasm	5.58×10^{-6}
Map	Map	Protein	Cytoplasm	3.72×10^{-4}
mat_tRNA_mRNA	$mat_{tRNA-mRNA}$	mRNA	Cytoplasm	3.35×10^{-5}

Table E.1 (Continued)

Id	Name	Type	Location	Initial Mass (pg)
Mat_tRNA	Mat_{tRNA}	Protein	Cytoplasm	2.23×10^{-3}
Met_tRNA	Met_{tRNA}	Amino-Acyl tRNA	Cytoplasm	1.19×10^{-4}
metK_mRNA	$metK_{mRNA}$	mRNA	Cytoplasm	5.58×10^{-6}
MetK	MetK	Protein	Cytoplasm	3.72×10^{-4}
metS_mRNA	$metS_{mRNA}$	mRNA	Cytoplasm	5.58×10^{-6}
MetS	MetS	Protein	Cytoplasm	3.72×10^{-4}
metT_mRNA	$metT_{mRNA}$	mRNA	Cytoplasm	1.67×10^{-5}
MetT	MetT	Protein	Cytoplasm	1.12×10^{-3}
Met	methionine	Amino Acid	Cytoplasm	6.38×10^{-5}
mgtA_mRNA	$mgtA_{mRNA}$	mRNA	Cytoplasm	5.58×10^{-6}
MgtA	MgtA	Protein	Cytoplasm	3.72×10^{-4}
mntH_mRNA	$mntH_{mRNA}$	mRNA	Cytoplasm	5.58×10^{-6}
MntH	MntH	Protein	Cytoplasm	3.72×10^{-4}
mraW_mRNA	$mraW_{mRNA}$	mRNA	Cytoplasm	5.58×10^{-6}
MraW	MraW	Protein	Cytoplasm	3.72×10^{-4}
NADH	NADH	Cofactor	Cytoplasm	6.38×10^{-5}
nadR_mRNA	$nadR_{mRNA}$	mRNA	Cytoplasm	5.58×10^{-6}
NadR	NadR	Protein	Cytoplasm	3.72×10^{-4}

Table E.1 (Continued)

Id	Name	Type	Location	Initial Mass (pg)
nadV_mRNA	$nadV_{mRNA}$	mRNA	Cytoplasm	5.58×10^{-6}
NadV	NadV	Protein	Cytoplasm	3.72×10^{-4}
NAD	NAD^{+}	Cofactor	Cytoplasm	6.38×10^{-5}
natT_mRNA	$natT_{mRNA}$	mRNA	Cytoplasm	2.23×10^{-5}
NatT	NatT	Protein	Cytoplasm	1.49×10^{-3}
ndk_mRNA	ndk_{mRNA}	mRNA	Cytoplasm	5.58×10^{-6}
Ndk	Ndk	Protein	Cytoplasm	3.72×10^{-4}
nhaB_mRNA	$nhaB_{mRNA}$	mRNA	Cytoplasm	5.58×10^{-6}
NhaB	NhaB	Protein	Cytoplasm	3.72×10^{-4}
nicotinamide	nicotinamide	Cofactor	Cytoplasm	1.02×10^{-7}
NMN	nicotinamide D-ribonucleotide	Cofactor	Cytoplasm	6.38×10^{-5}
nrdEF_mRNA	$nrdEF_{mRNA}$	mRNA	Cytoplasm	1.12×10^{-5}
NrdEF	NrdEF	Protein	Cytoplasm	7.43×10^{-4}
pal	fatty acids	Lipid	Cytoplasm	6.38×10^{-5}
pantetheine4P	pantetheine-4P	Cofactor	Cytoplasm	6.38×10^{-5}
pantothenate4P	pantothenate4P	Cofactor	Cytoplasm	6.38×10^{-5}
pantothenate	pantothenate	Cofactor	Cytoplasm	1.02×10^{-4}
PA	phosphatidate	Lipid	Cytoplasm	6.38×10^{-5}

Table E.1 (Continued)

Id	Name	Type	Location	Initial Mass (pg)
pdxY_mRNA	$pdxY_{mRNA}$	mRNA	Cytoplasm	5.58×10^{-6}
PdxY	PdxY	Protein	Cytoplasm	3.72×10^{-4}
PEP	phosphoenolpyruvate	Glycolytic	Cytoplasm	6.38×10^{-5}
PE	phosphatidylethanolamine	Lipid	Membrane	8.55×10^{-3}
pfkA_mRNA	$pfkA_{mRNA}$	mRNA	Cytoplasm	5.58×10^{-6}
PfkA	PfkA	Protein	Cytoplasm	3.72×10^{-4}
pgald	3-phosphoglyceraldehyde	Glycolytic	Cytoplasm	6.38×10^{-5}
pgi_mRNA	pgi_{mRNA}	mRNA	Cytoplasm	5.58×10^{-6}
Pgi	Pgi	Protein	Cytoplasm	3.72×10^{-4}
pgk_mRNA	pgk_{mRNA}	mRNA	Cytoplasm	5.58×10^{-6}
Pgk	Pgk	Protein	Cytoplasm	3.72×10^{-4}
Phe_tRNA	Phe_{tRNA}	Amino-Acyl tRNA	Cytoplasm	1.41×10^{-4}
pheS_mRNA	$pheS_{mRNA}$	mRNA	Cytoplasm	1.12×10^{-5}
PheS	PheS	Protein	Cytoplasm	7.43×10^{-4}
Phe	phenylalanine	Amino Acid	Cytoplasm	6.38×10^{-5}
pitA_mRNA	$pitA_{mRNA}$	mRNA	Cytoplasm	5.58×10^{-6}
PitA	PitA	Protein	Cytoplasm	3.72×10^{-4}
plsB_mRNA	$plsB_{mRNA}$	mRNA	Cytoplasm	5.58×10^{-6}

Table E.1 (Continued)

Id	Name	Type	Location	Initial Mass (pg)
PlsB	PlsB	Protein	Cytoplasm	3.72×10^{-4}
plsC_mRNA	$\text{plsC}_{\text{mRNA}}$	mRNA	Cytoplasm	5.58×10^{-6}
PlsC	PlsC	Protein	Cytoplasm	3.72×10^{-4}
pmf_mRNA	pmf_{mRNA}	mRNA	Cytoplasm	5.02×10^{-5}
Pmf	Pmf	Protein	Cytoplasm	3.35×10^{-3}
pol_RNA_mRNA	$\text{pol}_{\text{RNA}-\text{mRNA}}$	mRNA	Cytoplasm	4.46×10^{-5}
Pol_RNA	Pol_{RNA}	Protein	Cytoplasm	2.97×10^{-3}
ppa_mRNA	ppa_{mRNA}	mRNA	Cytoplasm	5.58×10^{-6}
Ppa	Ppa	Protein	Cytoplasm	3.72×10^{-4}
Pro_tRNA	Pro_{tRNA}	Amino-Acyl tRNA	Cytoplasm	1.24×10^{-4}
proS_mRNA	$\text{proS}_{\text{mRNA}}$	mRNA	Cytoplasm	5.58×10^{-6}
ProS	ProS	Protein	Cytoplasm	3.72×10^{-4}
prot_fold_mRNA	$\text{prot}_{\text{fold}-\text{mRNA}}$	mRNA	Cytoplasm	2.79×10^{-5}
Prot_fold	$\text{Prot}_{\text{fold}}$	Protein	Cytoplasm	1.86×10^{-3}
prot_transloc_mRNA	$\text{prot}_{\text{transloc}-\text{mRNA}}$	mRNA	Cytoplasm	2.79×10^{-5}
Prot_transloc	$\text{Prot}_{\text{transloc}}$	Protein	Cytoplasm	1.86×10^{-3}
Pro	proline	Amino Acid	Cytoplasm	6.38×10^{-5}
PRPP	5-phosphoribosyl-pyrophosphate	Pentose Phosphate	Cytoplasm	6.38×10^{-5}

Table E.1 (Continued)

Id	Name	Type	Location	Initial Mass (pg)
prsA_mRNA	$\text{prsA}_{\text{mRNA}}$	mRNA	Cytoplasm	5.58×10^{-6}
PrsA	PrsA	Protein	Cytoplasm	3.72×10^{-4}
psd_mRNA	psd_{mRNA}	mRNA	Cytoplasm	5.58×10^{-6}
Psd	Psd	Protein	Cytoplasm	3.72×10^{-4}
pssA_mRNA	$\text{pssA}_{\text{mRNA}}$	mRNA	Cytoplasm	5.58×10^{-6}
PssA	PssA	Protein	Cytoplasm	3.72×10^{-4}
PS	phosphatidylserine	Lipid	Cytoplasm	6.38×10^{-5}
ptsT_mRNA	$\text{ptsT}_{\text{mRNA}}$	mRNA	Cytoplasm	1.67×10^{-5}
PtsT	PtsT	Protein	Cytoplasm	1.12×10^{-3}
pykA_mRNA	$\text{pykA}_{\text{mRNA}}$	mRNA	Cytoplasm	5.58×10^{-6}
PykA	PykA	Protein	Cytoplasm	3.72×10^{-4}
pyrG_mRNA	$\text{pyrG}_{\text{mRNA}}$	mRNA	Cytoplasm	5.58×10^{-6}
PyrG	PyrG	Protein	Cytoplasm	3.72×10^{-4}
pyridoxalP	pyridoxalP	Cofactor	Cytoplasm	6.38×10^{-5}
pyridoxal	pyridoxal	Cofactor	Cytoplasm	1.02×10^{-4}
pyruvate	pyruvate	Glycolytic	Cytoplasm	6.38×10^{-5}
replisome_mRNA	$\text{replisome}_{\text{mRNA}}$	mRNA	Cytoplasm	5.58×10^{-5}
Replisome	Replisome	Protein	Cytoplasm	3.72×10^{-3}

Table E.1 (Continued)

Id	Name	Type	Location	Initial Mass (pg)
Rib_prot	protein in ribosomes	Protein	Cytoplasm	3.80×10^{-4}
Rib_RNA	rRNA in ribosomes	rRNA	Cytoplasm	2.76×10^{-2}
ribF_mRNA	$\text{ribF}_{\text{mRNA}}$	mRNA	Cytoplasm	5.58×10^{-6}
RibF	RibF	Protein	Cytoplasm	3.72×10^{-4}
ribM_mRNA	$\text{ribM}_{\text{mRNA}}$	mRNA	Cytoplasm	3.91×10^{-5}
RibM	RibM	Protein	Cytoplasm	2.60×10^{-3}
ribO_mRNA	$\text{ribO}_{\text{mRNA}}$	mRNA	Cytoplasm	2.79×10^{-4}
riboflavin	riboflavin	Cofactor	Cytoplasm	1.02×10^{-7}
ribose5P	ribose-5-P	Pentose Phosphate	Cytoplasm	6.38×10^{-5}
RibO	RibO	Protein	Cytoplasm	1.86×10^{-2}
ribulose5P	ribulose-5-P	Pentose Phosphate	Cytoplasm	6.38×10^{-5}
rpe_mRNA	rpe_{mRNA}	mRNA	Cytoplasm	5.58×10^{-6}
Rpe	Rpe	Protein	Cytoplasm	3.72×10^{-4}
rpiA_mRNA	$\text{rpiA}_{\text{mRNA}}$	mRNA	Cytoplasm	5.58×10^{-6}
RpiA	RpiA	Protein	Cytoplasm	3.72×10^{-4}
rti_RNA	rti_{RNA}	rRNA	Cytoplasm	1.67×10^{-5}
s7p	sedoheptulose-7P	Pentose Phosphate	Cytoplasm	6.38×10^{-5}
sahs	S-adenosyl-L-homocysteine	Cofactor	Cytoplasm	6.38×10^{-5}

Table E.1 (Continued)

Id	Name	Type	Location	Initial Mass (pg)
sam	S-adenosyl-L-methionine	Cofactor	Cytoplasm	6.38×10^{-5}
sbp	sedoheptulose-1,7-bisphosphate	Pentose Phosphate	Cytoplasm	6.38×10^{-5}
sept	septum	Lipid	Membrane	6.38×10^{-5}
Ser_tRNA	Ser_{tRNA}	Amino-Acyl tRNA	Cytoplasm	1.21×10^{-4}
serS_mRNA	$\text{serS}_{\text{mRNA}}$	mRNA	Cytoplasm	5.58×10^{-6}
SerS	SerS	Protein	Cytoplasm	3.72×10^{-4}
Ser	serine	Amino Acid	Cytoplasm	6.38×10^{-5}
sstT_mRNA	$\text{sstT}_{\text{mRNA}}$	mRNA	Cytoplasm	5.58×10^{-6}
SstT	SstT	Protein	Cytoplasm	3.72×10^{-4}
T_A2	A2 transporter	Protein	Membrane	4.75×10^{-4}
T_AG	AG transporter	Protein	Membrane	4.75×10^{-4}
T_Aro	Aro transporter	Protein	Membrane	4.75×10^{-4}
T_Bgt	Bgt transporter	Protein	Membrane	4.75×10^{-4}
T_Bzt	Bzt transporter	Protein	Membrane	4.75×10^{-4}
T_Cys	Cys transporter	Protein	Membrane	4.75×10^{-4}
T_FA	FA transporter	Protein	Membrane	4.75×10^{-4}
T_K	K transporter	Protein	Membrane	4.75×10^{-4}
T_lactate	lactate transporter	Protein	Membrane	4.75×10^{-4}

Table E.1 (Continued)

Id	Name	Type	Location	Initial Mass (pg)
T_Liv	Liv transporter	Protein	Membrane	4.75×10^{-4}
T_Met	Met transporter	Protein	Membrane	4.75×10^{-4}
T_Mg	Mg transporter	Protein	Membrane	4.75×10^{-4}
T_Mn	Mn transporter	Protein	Membrane	4.75×10^{-4}
T_Nat	Nat transporter	Protein	Membrane	4.75×10^{-4}
T_Na	Na transporter	Protein	Membrane	4.75×10^{-4}
T_Pi	Pi transporter	Protein	Membrane	4.75×10^{-4}
T_Sst	Sst transporter	Protein	Membrane	4.75×10^{-4}
T_U	U transporter	Protein	Membrane	4.75×10^{-4}
tcyP_mRNA	$\text{tcyP}_{\text{mRNA}}$	mRNA	Cytoplasm	5.58×10^{-6}
TcyP	TcyP	Protein	Cytoplasm	3.72×10^{-4}
THF510methylene	5,10-methylenetetrahydrofolate	Cofactor	Cytoplasm	6.38×10^{-5}
THF	tetrahydrofolate	Cofactor	Cytoplasm	6.38×10^{-5}
thiamine	thiamine	Cofactor	Cytoplasm	1.02×10^{-7}
Thr_tRNA	Thr_{tRNA}	Amino-Acyl tRNA	Cytoplasm	1.26×10^{-4}
thrS_mRNA	$\text{thrS}_{\text{mRNA}}$	mRNA	Cytoplasm	5.58×10^{-6}
ThrS	ThrS	Protein	Cytoplasm	3.72×10^{-4}
Thr	threonine	Amino Acid	Cytoplasm	6.38×10^{-5}

Table E.1 (Continued)

Id	Name	Type	Location	Initial Mass (pg)
thyA_mRNA	$thyA_{mRNA}$	mRNA	Cytoplasm	5.58×10^{-6}
ThyA	ThyA	Protein	Cytoplasm	3.72×10^{-4}
tkt_mRNA	tkt_{mRNA}	mRNA	Cytoplasm	5.58×10^{-6}
Tkt	Tkt	Protein	Cytoplasm	3.72×10^{-4}
tmk_mRNA	tmk_{mRNA}	mRNA	Cytoplasm	5.58×10^{-6}
Tmk	Tmk	Protein	Cytoplasm	3.72×10^{-4}
tpiA_mRNA	$tpiA_{mRNA}$	mRNA	Cytoplasm	5.58×10^{-6}
TpiA	TpiA	Protein	Cytoplasm	3.72×10^{-4}
TPP	thiamin diphosphate	Cofactor	Cytoplasm	6.38×10^{-5}
transF_mRNA	$transF_{mRNA}$	mRNA	Cytoplasm	6.69×10^{-5}
TransF	TransF	Protein	Cytoplasm	4.46×10^{-3}
tRNA_Ala	$tRNA_{Ala}$	tRNA	Cytoplasm	1.19×10^{-4}
tRNA_Arg	$tRNA_{Arg}$	tRNA	Cytoplasm	1.19×10^{-4}
tRNA_Asn	$tRNA_{Asn}$	tRNA	Cytoplasm	1.19×10^{-4}
tRNA_Asp	$tRNA_{Asp}$	tRNA	Cytoplasm	1.19×10^{-4}
tRNA_Cys	$tRNA_{Cys}$	tRNA	Cytoplasm	1.19×10^{-4}
tRNA_Gln	$tRNA_{Gln}$	tRNA	Cytoplasm	1.19×10^{-4}
tRNA_Glu	$tRNA_{Glu}$	tRNA	Cytoplasm	1.19×10^{-4}

Table E.1 (Continued)

Id	Name	Type	Location	Initial Mass (pg)
tRNA_Gly	$tRNA_{Gly}$	tRNA	Cytoplasm	1.19×10^{-4}
tRNA_His	$tRNA_{His}$	tRNA	Cytoplasm	1.19×10^{-4}
tRNA_Ile	$tRNA_{Ile}$	tRNA	Cytoplasm	1.19×10^{-4}
tRNA_Leu	$tRNA_{Leu}$	tRNA	Cytoplasm	1.19×10^{-4}
tRNA_Lys	$tRNA_{Lys}$	tRNA	Cytoplasm	1.19×10^{-4}
tRNA_Met	$tRNA_{Met}$	tRNA	Cytoplasm	1.19×10^{-4}
tRNA_Phe	$tRNA_{Phe}$	tRNA	Cytoplasm	1.19×10^{-4}
tRNA_Pro	$tRNA_{Pro}$	tRNA	Cytoplasm	1.19×10^{-4}
tRNA_Ser	$tRNA_{Ser}$	tRNA	Cytoplasm	1.19×10^{-4}
tRNA_Thr	$tRNA_{Thr}$	tRNA	Cytoplasm	1.19×10^{-4}
tRNA_Trp	$tRNA_{Trp}$	tRNA	Cytoplasm	1.19×10^{-4}
tRNA_Tyr	$tRNA_{Tyr}$	tRNA	Cytoplasm	1.19×10^{-4}
tRNA_Val	$tRNA_{Val}$	tRNA	Cytoplasm	1.19×10^{-4}
Trp_tRNA	Trp_{tRNA}	Amino-Acyl tRNA	Cytoplasm	1.37×10^{-4}
trpS_mRNA	$trpS_{mRNA}$	mRNA	Cytoplasm	5.58×10^{-6}
TrpS	TrpS	Protein	Cytoplasm	3.72×10^{-4}
Trp	tryptophan	Amino Acid	Cytoplasm	6.38×10^{-5}
trxA_mRNA	$trxA_{mRNA}$	mRNA	Cytoplasm	5.58×10^{-6}

Table E.1 (Continued)

Id	Name	Type	Location	Initial Mass (pg)
TrxAds	TrxAds	Protein	Cytoplasm	3.80×10^{-4}
TrxA	TrxA	Protein	Cytoplasm	3.72×10^{-4}
trxB_mRNA	$\text{trxB}_{\text{mRNA}}$	mRNA	Cytoplasm	5.58×10^{-6}
TrxB	TrxB	Protein	Cytoplasm	3.72×10^{-4}
Tyr_tRNA	Tyr_{tRNA}	Amino-Acyl tRNA	Cytoplasm	1.28×10^{-4}
tyrS_mRNA	$\text{tyrS}_{\text{mRNA}}$	mRNA	Cytoplasm	5.58×10^{-6}
TyrS	TyrS	Protein	Cytoplasm	3.72×10^{-4}
Tyr	tyrosine	Amino Acid	Cytoplasm	6.38×10^{-5}
UDP	uridine diphosphate	Nucleotide	Cytoplasm	6.38×10^{-5}
UMP	uridine monophosphate	Nucleotide	Cytoplasm	6.38×10^{-5}
upp_mRNA	upp_{mRNA}	mRNA	Cytoplasm	5.58×10^{-6}
Upp	Upp	Protein	Cytoplasm	3.72×10^{-4}
UTP	uridine triphosphate	Nucleotide	Cytoplasm	6.44×10^{-5}
Val_tRNA	Val_{tRNA}	Amino-Acyl tRNA	Cytoplasm	1.42×10^{-4}
valS_mRNA	$\text{valS}_{\text{mRNA}}$	mRNA	Cytoplasm	5.58×10^{-6}
ValS	ValS	Protein	Cytoplasm	3.72×10^{-4}
Val	valine	Amino Acid	Cytoplasm	6.38×10^{-5}
xyulose5P	xyulose-5-P	Pentose Phosphate	Cytoplasm	6.38×10^{-5}

Table E.1 (Continued)

Id	Name	Type	Location	Initial Mass (pg)
yloS_mRNA	$\text{yloS}_{\text{mRNA}}$	mRNA	Cytoplasm	5.58×10^{-6}
YloS	YloS	Protein	Cytoplasm	3.72×10^{-4}

REFERENCES

Neidhardt, H. E., Frederick C. Umbarger (1996). Chemical composition of *Escherichia coli*. *Escherichia coli and Salmonella Cellular and Molecular Biology*, 1, 13–16.

APPENDIX F

MINIMAL CELL MODEL EVENTS

Events describe instantaneous, discontinuous changes in the state of the model, and an implementation of events based on SBML is used in the MCM (Hucka et al., 2008). Because they cause discrete changes in the cell structure of behavior that occur instantaneously when the cell reaches some predefined condition, events require special mathematical treatment during a simulation. Detection of events also requires an algorithm that can detect when the firing of one event promotes another event to fire simultaneously (Nikolaev et al., 2006).

Table F.1 lists the 36 events in the Minimal Cell Model (MCM). Most of the events are associated with monitoring the limiting reagents of reactions with many substrates (e.g. protein synthesis). In Table F.1 these are identified as "min-switch" events.

Table F.1: Discrete physiological events in the Minimal Cell Model. The model has 36 events. "min-switch" events correspond to switches in limiting reactants for coarse-grained reactions that have many substrates (Section 4.18).

Event ID	Trigger
$\text{DNA}_{\text{initiation}}$	$(DnaG_{boundto\text{-}Ori} \geq init_{threshold}) \wedge (flag_{meth} == 1)$
$\text{DNA}_{\text{termination}}$	$(ForkPos_0 \geq 1.0)$
$\text{DNAp}_{\text{min-switch-to-dATP}}$	$(dATP < DNAp_{min})$
$\text{DNAp}_{\text{min-switch-to-dCTP}}$	$(dCTP < DNAp_{min})$
$\text{DNAp}_{\text{min-switch-to-dGTP}}$	$(dGTP < DNAp_{min})$
$\text{DNAp}_{\text{min-switch-to-dTTP}}$	$(dTTP < DNAp_{min})$
Division	$(SEP - sept_A \leq 0)$
$\text{DnaB}_{\text{active}}$	$\left(DnaB_{boundto\text{-}Ori} > 4.13482 \times 10^{-7}\right)$
$\text{DnaB}_{\text{inactive}}$	$\left(DnaB_{boundto\text{-}Ori} \leq 4.13482 \times 10^{-7}\right)$
$\text{HupA}_{\text{active}}$	$\left(HupA_{boundto\text{-}Ori} > 5.54452 \times 10^{-7}\right)$
$\text{HupA}_{\text{inactive}}$	$\left(HupA_{boundto\text{-}Ori} \leq 5.54452 \times 10^{-7}\right)$
$\text{M1p}_{\text{min-switch-to-Ala-tRNA}}$	$(Ala_{tRNA} < M1p_{min})$
$\text{M1p}_{\text{min-switch-to-Arg-tRNA}}$	$(Arg_{tRNA} < M1p_{min})$
$\text{M1p}_{\text{min-switch-to-Asn-tRNA}}$	$(Asn_{tRNA} < M1p_{min})$
$\text{M1p}_{\text{min-switch-to-Asp-tRNA}}$	$(Asp_{tRNA} < M1p_{min})$
$\text{M1p}_{\text{min-switch-to-Cys-tRNA}}$	$(Cys_{tRNA} < M1p_{min})$
$\text{M1p}_{\text{min-switch-to-Gln-tRNA}}$	$(Gln_{tRNA} < M1p_{min})$
$\text{M1p}_{\text{min-switch-to-Glu-tRNA}}$	$(Glu_{tRNA} < M1p_{min})$
$\text{M1p}_{\text{min-switch-to-Gly-tRNA}}$	$(Gly_{tRNA} < M1p_{min})$
$\text{M1p}_{\text{min-switch-to-His-tRNA}}$	$(His_{tRNA} < M1p_{min})$
$\text{M1p}_{\text{min-switch-to-Ile-tRNA}}$	$(Ile_{tRNA} < M1p_{min})$
$\text{M1p}_{\text{min-switch-to-Leu-tRNA}}$	$(Leu_{tRNA} < M1p_{min})$
$\text{M1p}_{\text{min-switch-to-Lys-tRNA}}$	$(Lys_{tRNA} < M1p_{min})$
$\text{M1p}_{\text{min-switch-to-Met-tRNA}}$	$(Met_{tRNA} < M1p_{min})$
$\text{M1p}_{\text{min-switch-to-Phe-tRNA}}$	$(Phe_{tRNA} < M1p_{min})$
$\text{M1p}_{\text{min-switch-to-Pro-tRNA}}$	$(Pro_{tRNA} < M1p_{min})$
$\text{M1p}_{\text{min-switch-to-Ser-tRNA}}$	$(Ser_{tRNA} < M1p_{min})$
$\text{M1p}_{\text{min-switch-to-Thr-tRNA}}$	$(Thr_{tRNA} < M1p_{min})$
$\text{M1p}_{\text{min-switch-to-Trp-tRNA}}$	$(Trp_{tRNA} < M1p_{min})$
$\text{M1p}_{\text{min-switch-to-Tyr-tRNA}}$	$(Tyr_{tRNA} < M1p_{min})$
$\text{M1p}_{\text{min-switch-to-Val-tRNA}}$	$(Val_{tRNA} < M1p_{min})$
$\text{MethState}_{\text{gt}-1}$	$(MethState > 1)$
$\text{RNAp}_{\text{min-switch-to-ATP}}$	$(ATP < RNAp_{min})$
$\text{RNAp}_{\text{min-switch-to-CTP}}$	$(CTP < RNAp_{min})$
$\text{RNAp}_{\text{min-switch-to-GTP}}$	$(GTP < RNAp_{min})$
$\text{RNAp}_{\text{min-switch-to-UTP}}$	$(UTP < RNAp_{min})$

REFERENCES

Hucka, M., Hoops, S., Keating, S., Le Novre, N., Sahle, S., et al. (2008). Systems biology markup language (SBML) level 2: Structures and facilities for model definitions. *Nature Precedings*. doi:doi.org/10.1038/npre.2008.2715.1.

Nikolaev, E., Atlas, J., and Shuler, M. L. (2006). Computer models of bacterial cells: from generalized coarse-grained to genome-specific modular models. *Journal of Physics: Conference Series*, 46, 322–326.

APPENDIX G

SENSITIVITY AND CONTROL ANALYSIS OF PERIODICALLY FORCED REACTION NETWORKS USING THE GREEN'S FUNCTION METHOD

The contents of this appendix are reproduced with permission from the *Journal of Theoretical Biology*[1]. This appendix contains the abstract of the paper. The full original paper was published by Nikolaev, Atlas, and Shuler (2007).

A general sensitivity and control analysis of periodically forced reaction networks with respect to small perturbations in arbitrary networks parameters and forcing frequency is presented using the Greens function method. A well-known property of sensitivity coefficients for periodic processes in dynamic systems is that the coefficients generally become unbounded as time tends to infinity. To circumvent the conceptual obstacle, a relative phase or fractional time variable is introduced so that when evaluated in terms of the new time variable, the periodic sensitivity coefficients can be calculated. By employing the Greens function method, the sensitivity coefficients can be defined using integral control operators that relate small perturbations in the networks parameters and forcing frequency to the variations in the metabolite concentrations and fluxes. The properties of such operators do not depend on a particular parameter-perturbation and are described by the summation and connectivity relationships within a control-matrix operator equation. The aim of the paper is to derive a general control-matrix operator equation for periodically forced reaction networks. To demonstrate the general method, the two limiting cases of high and low frequency are considered and an important case of

[1]Nikolaev, E.V., Atlas, J.C., and Shuler, M.L., 2007, "Sensitivity and control analysis of periodically forced reaction networks using the Greens function method", *Journal of Theoretical Biology*, vol. 247, pp. 442-461.

the simultaneous modulation of enzyme activities and external frequency is discussed. The developed framework is also illustrated by the calculation of the sensitivity and control coefficients for a simple two reaction pathway, where enzyme activities enter reaction rates linearly and specifically. We find that external force adds an important complicating factor as metabolic control can be continuously shifted between different groups of enzymes depending on the oscillatory phase. This shift can be controlled to some extent by the magnitude of the forcing frequency.

REFERENCES

Nikolaev, E. V., Atlas, J. C., and Shuler, M. L. (2007). Sensitivity and control analysis of periodically forced reaction networks using the Green's function method. *Journal of Theoretical Biology*, 247(3), 442–461. doi:10.1016/j.jtbi.2007.02.013.

APPENDIX H

SUPPLEMENT TO "INCORPORATING GENOME-WIDE DNA SEQUENCE INFORMATION INTO A DYNAMIC WHOLE-CELL MODEL OF *ESCHERICHIA COLI*: APPLICATION TO DNA REPLICATION"

The contents of this appendix are reproduced with permission from *IET Systems Biology*[1]. The information presented here is supplementary to Chapter 3 of this dissertation.

H.1 Dynamical Changes of DnaA-Binding Boxes Along the Replicating Chromosome

To obtain mathematical expressions for the number of DnaA-binding boxes described by formulas (3.3) - (3.5) in the main text, we consider first the simplest case, where a circular chromosome has only one moving $Fork_1$ with fractional position x_1 as shown in Figure 3.4. Let $y_1 = 1 - x_1$, y_1 is the fractional distance of $Fork_1$ from the DNA terminus. Then the total number of the DnaA-binding boxes with the cumulative number distribution function $F(y)$ along N_{chrom} synchronously replicating chromosomes can be calculated using the formulas (H.1) - (H.2).

$$S = N_{chrom} \cdot [F(1) + \Delta S^1] \tag{H.1}$$

[1]Atlas, J.C., Nikolaev, E.V., and Shuler, M.L., September 2008, "Incorporating Genome-Wide DNA Sequence Information into a Dynamic Whole-Cell Model of *Escherichia coli*: Application to DNA Replication", *IET Systems Biology*, vol. 2, no. 5, pp. 369-382,

$$\Delta S^1 = F(1) - F(y_1) \tag{H.2}$$

Here $F(1)$ is the total number of the DnaA boxes on the leading strand (see the main text), and ΔS^1 is the number of the DnaA boxes on the one newly synthesized lagging strand as shown in Figure 3.4(a). Using equation (H.2) in (H.1), we obtain (H.3). Using (3.2) from the main text with the omitted indices in (H.3), we can obtain (H.4).

$$S = N_{chrom} \cdot [2F(1) - F(y_1)] \tag{H.3}$$

$$S = N_{chrom} \cdot [2(a+b) - a \cdot y_1 - b \cdot y_1^2] \tag{H.4}$$

After simple algebraic manipulations, (H.4) can be transformed to (3.3) - (3.5), as in (H.5) at $y_2 = y_3 = 1$ corresponding to the absent $Fork_2$ and $Fork_3$. To check this we can use $y_2 = y_3 = 1$ in formulas (3.3) - (3.5) of the main text, leading to equation (H.5) which is equivalent to (H.4).

$$\begin{aligned} S &= N_{chrom} \cdot [2(a+b) - a \cdot y_1 - b \cdot y_1^2] \\ &= N_{chrom} \cdot [a \cdot (2 - y_1) + b \cdot (2 - y_1^2)] \\ &= N_{chrom} \cdot [a \cdot (y_1 + 2(1 - y_1)) + b \cdot (y_1^2 + 2(1 - y_1^2))] \end{aligned} \tag{H.5}$$

The cases with moving forks $Fork_2$ and $Fork_3$ can be considered in a similar way. Indeed, in the case when $Fork_1$ and $Fork_2$ are active, equation (H.1) can be rewritten in the form of equations (H.6) and (H.7).

$$S = N_{chrom} \cdot [F(1) + \Delta S^1 + 2\Delta S^2] \tag{H.6}$$

$$\Delta S^2 = F(1) - F(y_2) \tag{H.7}$$

Here N_{chrom}, $F(1)$, and ΔS^1 are defined as in (H.1) and (H.2), y_2 is the fractional distance of $Fork_2$ from the terminus of the replicating chromosome. $2\Delta S^2$ is the total number of the DnaA binding boxes within the *two* newly synthesized lagging strands as shown in Figure 3.4(b). Using (H.2) and (H.7) in (H.6), we obtain (H.8).

$$S = N_{chrom} \cdot [4F(1) - F(y_1) - 2F(y_2)] \tag{H.8}$$

Using (3.2) from the main text with the omitted indices in (H.8), we can obtain (H.9)

$$S = N_{chrom} \cdot [4(a+b) - (a \cdot y_1 + b \cdot y_1^2) - 2(a \cdot y_2 + b \cdot y_2^2)] \tag{H.9}$$

Similarly to (H.5), equation (H.9) can be rewritten in a form equivalent to (3.3) - (3.5) of the main text with $y_3 = 1$ corresponding to $Fork_3$ being absent.

Finally, when all three forks, $Fork_1$, $Fork_2$, and $Fork_3$ are active as shown in Figure 3.4(c), we obtain (H.10) and (H.11).

$$S = N_{chrom} \cdot [F(1) + \Delta S^1 + 2\Delta S^2 + 4\Delta S^3] \tag{H.10}$$

$$\Delta S^3 = F(1) - F(y_3) \tag{H.11}$$

Here y_3 is the fractional distance of $Fork_3$ from the terminus and $4\Delta S^3$ is the total number of the DnaA binding boxes within the *four* newly synthesized lagging strands as in Figure 3.4(c). Using (H.2), (H.7) and (H.11) in (H.10), we obtain (H.12).

$$S = N_{chrom} \cdot [8F(1) - F(y_1) - 2F(y_2) - 4F(y_3)] \tag{H.12}$$

Using (3.2) from the main text with the omitted indices in (H.12), we additionally obtain (H.13).

$$S = N_{chrom} \cdot [8(a+b) - (a \cdot y_1 + b \cdot y_1^2) - 2(a \cdot y_2 + b \cdot y_2^2) - 4(a \cdot y_3 + b \cdot y_3^2)] \tag{H.13}$$

It can be verified that (H.13) is equivalent to (3.3) - (3.5).

H.2 Ordered and Sequential Binding of DnaA-ATP Molecules to $oriC$

To calculate the discrete events corresponding to the formation of the replicon at $oriC$ and then its transitions between different states, we assume that about 28 DnaA-ATP molecules should bind to the replicon to begin the DNA replication process. Because there are *four* functional boxes in the $oriC$, R_1, R_2, R_3, and

R_4 (Margulies and Kaguni, 1996), there should presumably be *seven* ordered and sequential states that the replicon should pass through before all 28 DnaA-ATP molecules binds to the active replicon at $oriC$. These seven ordered and sequential states were experimentally observed for *E. coli* (Crooke et al., 1992; Margulies and Kaguni, 1996). Therefore in our approximation, we can assume that in average four DnaA-ATP molecules can bind to the replicon before its transition to the next state. It is also experimentally observed that DnaA-ATP molecules preferentially bind to the $oriC$ flanking functional boxes R_1 and R_4 with higher affinity relative to the central $oriC$ functional boxes R_2 and R_3. Therefore we can additionally postulate that there should presumably be a cooperative effect in the sense that the more DnaA-ATP molecules are titrated by the high-affinity H-boxes outside $oriC$ (i.e. the less H-boxes are available outside $oriC$), the more chance there is that next four DnaA-ATP molecules will bind to the replicon at $oriC$. Let S_{DnaA} be the number of H-boxes bound outside $oriC$ at time t, and let S_H be the total number of all H-boxes on the replicating chromosome at time t. We denote by N_B the number of the functional boxes in $oriC$, $N_B = 4$. Recall also that the binomial coefficient is defined as

$$\binom{n}{m} = \frac{n!}{m!(n-m)!}$$

where $n!$ is the factorial of n. To model the chance of the formation of the replicon at the "bare" $oriC$ at time t, we assume that four (i.e. N_B) free DnaA-ATP molecules can bind to S_H boxes giving rise to $\binom{S_H}{N_B}$ total possibilities. Additionally, we assume that four (i.e. N_B) of all bound H-boxes can be in $oriC$ at time t. This allows us to postulate the probability of the replicon formation corresponding to the case when four DnaA-ATP molecules bind to the bare $oriC$

at time t,

$$P_t \sim \binom{S_{DnaA}}{N_B} / \binom{S_H}{N_B} \tag{H.14}$$

Recall that $\Gamma(n+1) = n!$, when n is an integer (W.H. et al., 1988). Then (H.14) can be rewritten in the equivalent form (H.15).

$$P_t^0 \sim \frac{\Gamma(S_{DnaA}+1)}{\Gamma(S_{DnaA}-N_B+1)} \cdot \frac{\Gamma(S_H-N_B+1)}{\Gamma(S_H+1)} \tag{H.15}$$

which is more convenient for computations rather than direct calculation of factorials. We further assume that the replicon is formed at $oriC$ when the estimated P_t^0 is equal to the actual uniform probability (i.e. P_{oriC}) of the replicon "transition" defined in the main text (i.e., mathematically, the when algebraic event condition $P_t^0 = P_{oriC}$ is met).

To model the discrete transitions of the formed replicon between different states at $oriC$, we additionally assume that $\overline{S}_{DnaA}$ is the number of the bound H-boxes outside the formed replicon at time t and $\overline{S}_H$ is the number of free H-boxes. Similarly to (H.14), we can postulate the probability of the replicon transition between different states R, $R \in \{1, ..., 7\}$, at $oriC$ at time t

$$P_t \sim \binom{\overline{S}_{DnaA}}{N_B} / \binom{\overline{S}_H}{N_B} \tag{H.16}$$

Using identity $\Gamma(n+1) = n!$, (H.16) can be rewritten in a more computationally convenient form (H.17)

$$P_t \sim \frac{\Gamma(\overline{S}_{DnaA} + 1)}{\Gamma(\overline{S}_{DnaA} - N_B + 1)} \cdot \frac{\Gamma(\overline{S}_H - N_B + 1)}{\Gamma(\overline{S}_H + 1)} \tag{H.17}$$

Again, we assume that the replicon transition happens when the estimated P_t is equal to the actual uniform probability (i.e. P_{oriC}) of the replicon transition as discussed in the main text (i.e., mathematically, the corresponding algebraic discrete event condition is $P_t = P_{oriC}$).

References

Crooke, E., Castuma, C. E., and Kornberg, A. (1992). The chromosome origin of *Escherichia coli* stabilizes DnaA protein during rejuvenation by phospholipids. *Journal of Biological Chemistry*, 267(24), 16779–16782.

Margulies, C. and Kaguni, J. M. (1996). Ordered and sequential binding of DnaA protein to *oriC*, the chromosomal origin of *Escherichia coli*. *Journal of Biological Chemistry*, 271(29), 17035–17040.

W.H., P., B.P., F., and Teukolsky S.A., V. W. (1988). *Numerical Recipes in C*, chapter 6.1, pages 213–216. Cambridge University Press.

APPENDIX I

SUPPLEMENTAL WEBSITE

A projected long-term impact of this dissertation is to make the Minimal Cell Model (MCM) available to a wide audience. The model is available in the Systems Biology Markup Language (SBML) (Hucka et al., 2003, 2008) with model a simulator called SloppyCell available in Python (Gutenkunst et al., 2007a). The MCM makes heavy use of SloppyCell's simulation features, but it was not possible to use the parameter estimation and sensitivity analysis features of SloppyCell with the MCM because of its large size. Having the MCM available in SBML, however, means that a researcher could potentially simulate the system using any simulation package that accepts SBML input.

To establish a complete record of the work presented here, we have registered a supplemental website for the MCM project at `http://minimalcell.bme.cornell.edu` that will include all the computer code used in this dissertation.

The program code will be available in a distribution archive that contains the following top-level directories:

- `Data` contains the Microsoft Access 2003 database that defines all the compartments, species, and genes in the MCM (`mcm_db.mdb`). It also includes data related to calculating the initial conditions for the model (`InitialConditions.py`) and function definitions for importing information from the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000) into the Access database.
- `Documentation` contains files related to creating the automatically

generated model documentation. There is a documentation template file (`documentation_template.tex`) which contains the base structure for generating the model documents. To actually generate the model documents, use the script `quick_documents.py` in the `Run` directory.

- `KEGG` contains a file, `KEGGInterface.py` which defines functions for making a connection to the KEGG database online using the Simple Object Access Protocol (SOAP).
- The `lpsolve` directory contains a wrapper class for the lpsolve open source mix-ed integer linear program software, used under the LGPL (Berkelaar et al., 2010). lpsolve is used to estimate rate constants for the MCM. Only the wrapper class is included in the project distribution. The software is not included with the MCM distribution, and it must be installed separately. lpsolve is available from `http://lpsolve.sourceforge.net/`.
- `MCM_base` defines the modules, or sub-models, of the MCM. For example, `Compartments.py`, `Reactions.py`, and `Species.py` modules instantiate data structures related to the compartments, reactions, and species used in the model.
- `MCM_structures` contains all the class definitions for modeling structures used in the MCM, such as Reactions and Parameters. These modeling structures are described in Chapter 4.
- `Run` is where all scripts related to model simulation and experimentation are stored. Each file has a comment preamble that explains its purpose. Those just starting with the MCM should try to run the `run_base.py` and `run_base_analysis.py` scripts.

- `Sequence` contains functions for manipulating gene and protein sequences. For example, it includes functions for automatically converting an amino acid sequence to a protein synthesis stoichiometry.
- `Testing` contains a Python testing suite built using the `unittest` framework provided in Python. The format of this testing suite was modeled after the testing suite used in SloppyCell (Gutenkunst et al., 2007a).

The MCM code makes heavy use of SloppyCell, a Python software package for simulation and analysis of biomolecular networks (Gutenkunst et al., 2007a). SloppyCell has been applied to several biological systems of interest (Waterfall et al., 2006; Gutenkunst et al., 2007b,c). Significant updates have been made to SloppyCell as part of the current research to adapt it to simulating a model of this size and complexity. Support was added for several previously unsupported features of the SBML specification, including algebraic rules, model constraints, and event trigger functions with logical expressions. The current generation of the MCM has only been tested on Windows XP using SloppyCell built directly from the Concurrent Version System (CVS) source. The SloppyCell source is available at `http://sloppycell.sourceforge.net/`, and the MCM website will include up-to-date instructions for simulating the MCM reaction network to work with SloppyCell on Windows XP. The current version of those instructions are summarized in Section I.1. Sections I.2 and I.3 show short examples of usage for the MCM software. More extensive examples will be posted at the MCM website at `http://minimalcellmodel.bme.cornell.edu`.

I.1 Installing the Minimal Cell Model on Windows XP

Installing and simulating the MCM software has several requirements. We have only worked with the MCM on Windows XP, but because Python is platform-independent it should be possible to install the software under alternate operating systems. Some features, however, do require Windows (e.g., reconstructing the model from the start using the Microsoft Access database of compartments, genes, and chemical species). The following steps outline the procedure that we currently use to install SloppyCell and the MCM under Windows XP.

1. Install Python (version 2.6.4 recommended)
    - Add the root directory of the python installation to the `PATH` environment variable (e.g. `C:\Python26`).
2. Install `libSBML`.
    - Use `libSBML 4.0` compiled with `vc90` or later.
    - Set the `PYTHONPATH` environment variable to the Python bindings directory of the `libSBML` installation.
3. Install `scipy` and `numpy`. Make sure the installation matches the version of Python installed.
    - Use `scipy-0.7.1` or later.
    - Use `numpy-1.3.0` or later.
    - Note: `scipy 0.7.1` and `numpy 1.4.0` are incompatible.
4. Install `matplotlib-0.99` or later for the appropriate version of Python.

5. Install MinGW, making sure to include the `g77`, `g++`, and `MinGWMake` compiler apps when prompted.

    - add `C:\MinGW\bin` to the `PATH` environment variable.

6. If SloppyCell will be installed from the CVS source of the latest version, install TortoiseCVS client.

    - configure an SSH key using the instructions at `http://sourceforge.net/apps/trac/sourceforge/wiki/SSH%20keys`.
    - configure TortoiseCVS the first time you do a checkout using the instructions at `http://sourceforge.net/apps/trac/sourceforge/wiki/TortoiseCVS%20instructions`.

7. Install the Microsoft .NET framework (this will be required for the Visual C++ compilers).
8. Install Microsoft Visual C++ Express version. This is the C compiler we use for building SloppyCell.
9. Install SloppyCell from `http://sloppycell.sourceforge.net/`.

    - The SloppyCell website has an installer available for Win32 platforms.
    - If you download the source from the CVS server, unzip the downloaded files into the site-packages directly of your Python installation.
    - at the command prompt in the SloppyCell directory, run

      `"python setup.py build -cmsvc install --install-lib=..\."`.

- To test the SloppyCell installation, navigate to the `SloppyCell\Testing` directory and execute the `test.py` script from the command line.

10. Install `pyodbc` to be able to connect to the Microsoft Access database of compartments, genes, and chemical species (this is required to regenerate the MCM rather than using a precompiled SBML file).
11. Install `lpsolve 5.5` (`http://lpsolve.sourceforge.net/5.5/`) so that the system can calculate rate constants (Berkelaar et al., 2010).
    - You may need to copy the `lpsolve55.dll` file from the `extra\Python` directory to somewhere in the system path.
    - Ensure that the `lpsolve` directory was included in your distribution of the MCM.
12. (Optional) Install Graphviz (`http://www.graphviz.org/`), and add the Graphviz bin directory to PATH environment variable so that it can be called from command line (e.g. `C:\Program Files\Graphviz2.27\bin`).
13. (Optional) Install `SoapPy` for KEGG Application Programming Interface (API) access.
    - This also requires the Python `fpconst` module (`http://pypi.python.org/pypi/fpconst/0.7.2`). The KEGG wrapper is known to work with `fpconst 0.7.2`.
    - To get `fpconst` working, you may need to move "from future" imports in `Client.py`, `Types.py`, and `Server.py`.
14. (Optional) Install Processing 1.0.9 or later for cell growth visualization (`http://processing.org`).

15. (Optional) Install QuickTime to be able to save cell growth movies generated by Processing (`http://www.apple.com/quicktime/`).

16. (Optional) Install Circos to the `C:\Apps\Circos` directory so that the circos perl script is at `C:\Apps\Circos\bin\circos` (`http://mkweb.bcgsc.ca/circos/`).

    - Requires Perl 5.8x or newer. ActivePerl 5.10 is recommended (`http://www.activestate.com/activeperl/`).
    - PERL Packages installed from ActivePerl's package manager:
      - `Clone`
      - `Config::General`
      - `Math::Bezier`
      - `Math::Round`
      - `Math::VecStat`
      - `Params::Validate`
      - `Readonly`
      - `Set::IntSpan`
      - `Statistics::Descriptive`

17. Obtain the MCM distribution archive from `http://minimalcellmodel.bme.cornell.edu`. Unpack the archive into some local directory.

    - To quickly see if your installation is working, navigate to the `MCM\Run` directory, and from the command line execute the `run_base.py` and `run_base_analysis.py` scripts. `run_base.py` just initializes a Cell object. `run_base_analysis.py` initializes the cell object and performs a simulation from the default condition,

and then saves output from the simulation to the `MCM\Run\figs` directory.

- More extensive tests of the MCM installation are in the `MCM\Testing` directory. To execute the Python tests, navigate to the `MCM\Testing` directory and run the `test.py` file from the command line. Note: Depending on your computer hardware, running the full test suite may take over 24 hours.

I.2 Simulation and Integration

The following example shows how to load and simulate the default MCM. The model cell object is loaded and initialized, and then a SloppyCell reaction network is generated. This reaction network is integrated in time. This sort of integration is the basis for all the computational experiments that are performed with the MCM, so understanding it is essential to progressing to more complicated examples.

```
1  ############################################################
2  #
3  # This listing demonstrates how to load the Minimal Cell
4  # Model 'cell' object and then do a time integration of the
5  # reaction network using SloppyCell.
6  #
7  ############################################################
8
9  # The cell object is defeined in Components.py
10 from MCM_base.Components import *
```

```
11
12 # The ReactionNetworks directory of SloppyCell contains
13 # modules that handle reaction networks and time-integration
14 from SloppyCell.ReactionNetworks import *
15
16 # For the new cell, we must calculate reaction rate
17 # constants
18 cell.calculate_initial_rates()
19
20 # We also set constraints on the cell so that SloppyCell
21 # will know that if any species obtains a mass < 0 that
22 # the model simulation has become invalid.
23 cell.set_constraints()
24
25 # Construct a SloppyCell network.
26 cell.construct_ss_net()
27 net = cell.net
28
29 # Setting the network parameters to being non-optimizable
30 # will speed network compilation. Similarly, we disabale
31 # the compilation of the network's derivative functions.
32 for par in net.GetParameters().keys():
33     net.set_var_optimizable(par, is_optimizable=False)
34 net.disable_deriv_funcs()
35
36 # compiling the network is the final step before
37 # integration
38 net.compile()
39
40 # We define a time range for integration, and then call
41 # the integrate function of SloppyCell's 'Dynamics' module
42 times = scipy.linspace(0, 10, 500)
```

```
43 traj = Dynamics.integrate(net, times, fill_traj=False,
44                            return_derivs=True,
45                            redirect_msgs=False)
46
47 # The traj variable contains the results of the simulation,
48 # which can be plotted using SloppyCell or analyzed as the
49 # current experiment demands.
```

I.3 Computational Experiments

The Experiment class provides structures and functions to assist in simulating the model over a range of parameter values. This is useful when, for example, one wants to demonstrate the effect of changing a particular rate constant on the model's overall behavior. Each Experiment receives as input a list of conditions that will be tested when the Experiment is 'run'. The MCM website will list more complicated examples, but this listing shows a basic experiment where all the rate constants in the model are scaled simultaneously by a range of factors.

```
1
2 ############################################################
3 #
4 # This listing demonstrates how to run a simple
5 # computational experiment using the Minimal Cell Model.
6 #
7 ############################################################
8
9 # The cell object is defeined in Components.py
10 from MCM_base.Components import *
```

```
11
12 # The following imports load functions and classes related
13 # to experiments
14 from MCM_structures.Experiment_mod import *
15 from MCM_structures.event_manip import *
16
17 # We import the SloppyCell integration modules and rename
18 # the module as SRN for convenience.
19 import SloppyCell.ReactionNetworks as SRN
20
21 # Initial cell preperation and reaction network generation
22 cell.calculate_initial_rates()
23 cell.set_constraints()
24 cell.construct_ss_net()
25 net = cell.net
26
27 # We create a copy of the reaction network so that the
28 # original is not modified during the experiment
29 net_exper = net.copy()
30
31 # An Experiment object accepts a list of conditions
32 # that the Experiment what parameter values to use for
33 # each data point or trial.
34
35 # This experiments varies all of the reaction rate constants
36 # in the model (vms) by some scale. We selecte a range of 20
37 # scale values evenly spaced on a log scale.
38 scales = scipy.logspace(scipy.log10(0.01),scipy.log10(10), 20)
39
40
41 # Each entry in the condition list will be a dictionary that
42 # maps parameter names to values
```

```
conditions = []

for scale in scales:
    scale_condition = {}
    for r in cell.reactions.values():
        (vm, val) = r.v_m
        scale_condition[vm] = val*scale
    conditions.append(scale_condition)


# Chose a default integration time. The Experiment object
# will automatically increase the integration time if
# necessary to find a steady-state for the simulation.
times = scipy.linspace(0,15,2)
exper = Experiment('vm_scaling_experiment', cell,
                   net_exper, conditions, times)

# Run the experiment for all the conditions specified
exper.run()

# Plot results from the experiment
exper.plot_single_values('mu_g', plot_type='division',
                         vs='scaled', xs_alt=scales,
                         xlabel_alt ='scalefactor')

# Save the results of the experiment. This is useful
# because running the entire experiment can take a long
# time. Saving the results can allow us to quickly revisit
# old experiments without starting over.
exper.save('%s.pickle'%(exper.id))
```

REFERENCES

Berkelaar, M., Eikland, K., and Notebaert, P. (2010). lpsolve - Open source (mixed-integer) linear programming system, version 5.1.0.0, `http://lpsolve.sourceforge.net/`.

Gutenkunst, R. N., Atlas, J. C., Casey, F. P., Kuczenski, R. S., Waterfall, J. J., et al. (2007a). SloppyCell, `http://sloppycell.sourceforge.net/`.

Gutenkunst, R. N., Casey, F. P., Waterfall, J. J., Myers, C. R., and Sethna, J. P. (2007b). Extracting falsifiable predictions from sloppy models. *Ann N Y Acad Sci*, 1115, 203–211. doi:10.1196/annals.1407.003.

Gutenkunst, R. N., Waterfall, J. J., Casey, F. P., Brown, K. S., Myers, C. R., et al. (2007c). Universally sloppy parameter sensitivities in systems biology models. *PLoS Comput Biol*, 3(10), 1871–1878. doi:10.1371/journal.pcbi.0030189.

Hucka, M., Finney, A., Sauro, H. M., Bolouri, H., Doyle, J. C., et al. (2003). The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4), 524–531.

Hucka, M., Hoops, S., Keating, S., Le Novre, N., Sahle, S., et al. (2008). Systems biology markup language (SBML) level 2: Structures and facilities for model definitions. *Nature Precedings*. doi:doi.org/10.1038/npre.2008.2715.1.

Kanehisa, M. and Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1), 27–30.

Waterfall, J. J., Casey, F. P., Gutenkunst, R. N., Brown, K. S., Myers, C. R., et al. (2006). Sloppy-model universality class and the vandermonde matrix. *Phys Rev Lett*, 97(15), 150601.